

Optimal control of elliptic PDEs at points

CHARLES BRETT[†], ANDREAS DEDNER[‡] AND CHARLES ELLIOTT[§]
Department of Mathematics, University of Warwick, Coventry, CV4 7AL, UK

[November 19, 2014]

We consider an elliptic optimal control problem where the objective functional contains evaluations of the state at a finite number of points. In particular, we use a fidelity term that encourages the state to take certain values at these points, which means our problem is related to ones with state constraints at points. The analysis and numerical analysis differs from when the fidelity is in the L^2 norm because we need the state space to embed into the space of continuous functions. In this paper we discretise the problem using two different piecewise linear finite element methods. For each discretisation we use two different approaches to prove a priori L^2 error estimates for the control. We discuss the differences between these methods and approaches and present numerical results that agree with our analytical results.

Keywords: elliptic optimal control problem; point evaluations; finite element method; error estimates

1. Introduction

In this paper we study an elliptic optimal control problem with an objective functional containing the distance between the state and prescribed values at a finite number of prescribed points. This contrasts with standard elliptic optimal control problems, where typically the objective functional contains the L^2 distance between the state and the desired state over the whole domain. So for a bounded domain $\Omega \subset \mathbb{R}^n$ ($n = 2$ or 3) with boundary $\partial\Omega$ we consider the problem:

$$\min \frac{1}{2} \sum_{\omega \in I} (y(\omega) - g_\omega)^2 + \frac{\nu}{2} \|\eta\|_{L^2(\Omega)}^2$$

subject to the state equation

$$\begin{aligned} Ay &= \eta & \text{in } \Omega \\ y &= 0 & \text{on } \partial\Omega \end{aligned} \tag{1.1}$$

and the control constraints

$$a \leq \eta \leq b.$$

Here $I \subset \Omega$ is a finite set of points, $\{g_\omega\}_{\omega \in I} \subset \mathbb{R}$ are prescribed values at these points, $\nu > 0$ is the cost of control, A is an elliptic operator, and $a, b \in \mathbb{R}$ with $a < b$ are lower and upper bounds for the control. We give the precise statement of the problem using function spaces in Section 3.

The motivation for the point fidelity term is that in some applications we may only care about the state being close to given values at certain points in the domain. Controlling the state using a distributed

[†]Corresponding author. Email: ceabrett@gmail.com. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Grant EP/H023364/1.

[‡]Email: a.s.dedner@warwick.ac.uk

[§]Email: c.m.elliott@warwick.ac.uk

norm over the whole domain yields weaker control at points. The point fidelity term encourages the state to take certain values at points, so our problem is closely related to one which imposes hard constraints on the state at points. Imposing hard state constraints can often lead to an optimal control with a very high cost, whereas our point fidelity term allows for a compromise between how close the state is to the prescribed values and the cost of the control. On the other hand, we will prove later that as we increase the relative weighting given to the point fidelity term, the solutions of point control problems converge weakly to the solution of a problem with point state constraints.

In the literature there are computational results for PDE optimal control problems with objective functionals that contain point evaluations of the state. However we have not found any literature that contains a numerical analysis of such problems. The book Tröltzsch (2010) formulates an optimal control problem where the objective functional is the state evaluated at a point, but does not discuss numerical methods for solving it. The paper Unger & Tröltzsch (2001) considers optimally controlling the cooling of steel. This problem is formulated with an objective functional that contains the temperature of the steel at a number of points (i.e. point evaluations of the state) as this makes the problem more tractable. The paper focuses on computational results and the numerical analysis is not considered. The medical imaging problem of electrical impedance tomography (see e.g. Hintermüller & Laurain (2008)) could be formulated as an inverse problem with a point fidelity term (but with the points on the boundary). This is because one reconstructs a conductivity based on measurements of the voltage over small regions, which could be approximated by measurements at points. In the paper Brett *et al.* (2013) (written by ourselves) the point fidelity term is used for the optimal control of elliptic variational inequalities. The difficulty of the nonlinear control-to-state operator means that an a posteriori error estimator is derived but a priori error estimates are not considered.

Our aim is to fill a gap in the literature by studying in detail the numerical analysis of a finite element approximation of the above point control problem, which could be considered the canonical optimal control problem with an objective functional containing point evaluations of the state. However related problems have been considered in the literature. The recent paper Gong *et al.* (2014) considers elliptic optimal control problems with controls at points and on other lower dimensional manifolds. The numerical analysis of these problems leads to mathematical difficulties similar to those in this paper. In particular, when the control is at points the state equation has delta functions on the right hand side, whereas in our problem the adjoint equation has delta functions. In both cases this means low regularity of the state/adjoint. In the paper Brett *et al.* (2014) and thesis Brett (2014) theory is developed for an elliptic optimal control problem where the fidelity term is an integral along a surface of codimension 1, which is also a set of measure zero relative to the domain. In papers such as Casas *et al.* (2012) and Pieper & Vexler (2013) elliptic optimal control problems are considered where the control spaces are spaces of measures.

Regularity issues are also faced by elliptic optimal control problems with state constraints. The paper Leykekhman *et al.* (2013) proves error estimates for problems with state constraints at a finite number of points. Note that this paper also proves improved error estimates for graded triangulations (such triangulations are locally refined towards the singularities but have asymptotically the same number of elements for a given triangulation size), but we do not consider these. The paper Deckelnick & Hinze (2007) proves error estimates for the case of global (as opposed to point) state constraints, but for a state equation with Neumann boundary conditions. Parabolic optimal control problems often contain point evaluations in time of the state, but these are functions over the space domain and the technicalities of the numerical analysis are different. A review of the analysis for standard elliptic and parabolic optimal control problems can be found in Tröltzsch (2010) and a review of the numerical analysis can be found in Hinze *et al.* (2009).

TABLE 1. *The main a priori error estimates proved for $\|u - u_h\|_{L^2(\Omega)}$.*

Discretisation	(M1 _h)	(M1 _h) = (M2 _h)	(M2 _h)
Dimensions	$n = 2$	$n = 2, 3$	$n = 2, 3$
Constraints	both	$b = -a = \infty$	both
Approach 1	$O(h)$	$O(h^{2-\frac{n}{2}})$	$O(h^{2-\frac{n}{2}})$
Approach 2	-	$O(h^{2-\frac{n}{2}-\varepsilon})$	$O(h^{2-\frac{n}{2}-\varepsilon})$
Numerics	-	$O(h^{2-\frac{n}{2}})$	$O(h^{2-\frac{n}{2}})$

In this paper we use two different methods of discretising our problem with finite elements. The first method is to explicitly discretise the control by minimising over a space of discrete controls, leading to discrete problem (M1_h) (see (4.10)). The second method is to implicitly discretise the control through a discrete control-to-state operator using the variational discretisation concept from Hinze (2005), leading to discrete problem (M2_h) (see (4.13)). We later observe that when there are no control constraints these two methods may lead to equivalent discrete problems. We are not able to prove an estimate for (M1_h) in dimension 3 with control constraints, which motivates us to use (M2_h) for our implementation despite it being less standard to solve computationally.

Next we use two different approaches to prove a priori error estimates for the $L^2(\Omega)$ error in the control for these discrete problems. The first approach (Approach 1, Section 5.1) is inspired by the paper Casas & Tröltzsch (2003) and the second approach (Approach 2, Section 5.2) is inspired by the paper Deckelnick & Hinze (2007). The main estimates we prove are summarised in Table 1, where $\varepsilon > 0$ is arbitrary. We see that Approach 2 does not offer any better error estimates than Approach 1. However we include Approach 2 because it is simpler when it applies. Numerical results confirm that the error estimates are realised for (M2_h).

In the next section we introduce some notation. In Section 3 we formulate the optimal control problem precisely and prove some analytical results. In Section 4 we discretise using the finite element method. In Section 5 we prove a priori error estimates for the L^2 error in the control. In Section 6 we show numerical results.

2. Notation

We begin by introducing some function spaces that are needed to formulate the optimal control problem precisely.

Let the domain $\Omega \subset \mathbb{R}^n$ ($n = 2$ or 3) be a bounded open set that either has a $C^{1,1}$ boundary or is convex with a polygonal (for $n = 2$) or polyhedral (for $n = 3$) boundary. Both $C(\bar{\Omega})$ and its subspace $C_0(\Omega)$ (of functions that are zero on $\partial\Omega$) are Banach spaces when endowed with the supremum norm, $\|\cdot\|_\infty$. For $n = 2$ or 3 the Sobolev space $H^2(\Omega)$ is continuously embedded into $C(\bar{\Omega})$ (see e.g. Adams & Fournier (2003)), so $H^2(\Omega) \cap H_0^1(\Omega) \subset C_0(\Omega)$. By different versions of the Riesz Representation Theorem (see e.g. Theorems 2.14 and 6.19 in Rudin (1987)) the dual spaces of $C(\bar{\Omega})$ and $C_0(\Omega)$ can both be identified with the space $\mathcal{M}(\Omega)$ of real regular Borel measures on Ω . In particular, for $\mu \in \mathcal{M}(\Omega)$ and $v \in C(\bar{\Omega})$ define the duality pairing

$$\langle \mu, v \rangle_{\mathcal{M}(\Omega)} := \int_{\Omega} v d\mu,$$

where the integral is the Lebesgue integral with respect to μ . Here $\langle \mu, v \rangle_{\mathcal{M}(\Omega)}$ abbreviates $\langle \mu, v \rangle_{\mathcal{M}(\Omega), C(\bar{\Omega})}$.

Then for each $z \in C(\bar{\Omega})^*$ there exists a unique $\mu \in \mathcal{M}(\Omega)$ such that

$$z(v) = \langle \mu, v \rangle_{\mathcal{M}(\Omega)} \quad \forall v \in C(\bar{\Omega}). \quad (2.1)$$

The same result holds for $z \in C_0(\Omega)^*$ using the same definition of $\langle \mu, v \rangle_{\mathcal{M}(\Omega)}$ but with $v \in C_0(\Omega)$. We prefer to write $\int_{\Omega} v d\mu$ but will sometimes use $\langle \mu, v \rangle_{\mathcal{M}(\Omega)}$ to simplify notation. Note that $\mathcal{M}(\Omega)$ is a Banach space with the norm

$$\|\mu\|_{\mathcal{M}(\Omega)} := |\mu|(\Omega) = \sup \left\{ \int_{\Omega} v d\mu : v \in C_0(\Omega) \text{ and } \|v\|_{\infty} \leq 1 \right\},$$

where $|\mu|$ is called the total variation of μ . For example, the Dirac measure centred at a point $\omega \in \Omega$, which we denote by δ_{ω} , is contained in $\mathcal{M}(\Omega)$ and $\|\delta_{\omega}\|_{\mathcal{M}(\Omega)} = 1$.

We will need the following embedding results for the Sobolev spaces $W^{1,s}(\Omega)$, where $V \hookrightarrow W$ denotes that V is continuously embedded into W .

REMARK 2.1 From Adams & Fournier (2003) we have that:

- For $s > n$, $W^{1,s}(\Omega) \hookrightarrow C(\bar{\Omega})$;
- For $s > \frac{2n}{n+2}$, $W^{1,s}(\Omega) \hookrightarrow L^2(\Omega)$;
- For $s < \frac{2n}{n-2}$, $H^2(\Omega) \hookrightarrow W^{1,s}(\Omega)$.

Consider the Dirichlet problem (1.1), where the differential operator A acting on a function $z : \Omega \rightarrow \mathbb{R}$ is defined by

$$Az = - \sum_{i,j=1}^n \partial_{x_j} (a_{ij} \partial_{x_i} z) + a_0 z$$

with

$$\begin{aligned} a_0 &\in L^{\infty}(\Omega), \quad a_0(x) \geq 0 \quad \text{for a.e. } x \in \Omega, \\ a_{ij} &= a_{ji} \in C^{0,1}(\bar{\Omega}), \\ \exists \alpha > 0 \text{ s.t. } \sum_{i,j=1}^n a_{ij}(x) \xi_i \xi_j &\geq \alpha |\xi|^2, \quad \forall x \in \Omega, \xi \in \mathbb{R}^n. \end{aligned}$$

In particular, $A = -\Delta$ satisfies these assumptions. We want to work with a weak formulation of (1.1). Define the conjugate q' of q to be the real number such that $\frac{1}{q} + \frac{1}{q'} = 1$, and define the bilinear form $a : W_0^{1,q}(\Omega) \times W_0^{1,q'}(\Omega) \rightarrow \mathbb{R}$ associated to A by

$$a(z, v) = \sum_{i,j=1}^n \int_{\Omega} a_{ij}(x) \partial_{x_i} z(x) \partial_{x_j} v(x) dx + \int_{\Omega} a_0(x) z(x) v(x) dx,$$

where the derivatives are taken in the weak sense. By a standard result, for $\eta \in L^2(\Omega)$ there is a unique $y \in H_0^1(\Omega)$ satisfying

$$a(y, v) = (\eta, v) \quad \forall v \in H_0^1(\Omega). \quad (2.2)$$

Here and throughout this paper (\cdot, \cdot) denotes the $L^2(\Omega)$ inner product. With our assumptions on the domain Ω we have that $y \in H^2(\Omega) \cap H_0^1(\Omega)$ and

$$\|y\|_{H^2(\Omega)} \leq C \|\eta\|_{L^2(\Omega)}.$$

Here and throughout this paper C is a positive constant that may vary from line to line and is independent of the variables it precedes (e.g. in the above equation C is independent of η). For a proof of this regularity and stability result see Theorems 2.2.2.3 and 3.2.1.2 in Grisvard (1985). Since $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ we in fact have that $y \in C_0(\Omega)$ and

$$\|y\|_\infty \leq C\|\eta\|_{L^2(\Omega)}. \quad (2.3)$$

We define the control-to-state operator $S : L^2(\Omega) \rightarrow C_0(\Omega)$ to map $\eta \in L^2(\Omega)$ to the solution $y \in C_0(\Omega)$ of (2.2). S is linear, and also continuous by (2.3), so S has an adjoint operator. Using (2.1) we can define the adjoint $S^* : \mathcal{M}(\Omega) \rightarrow L^2(\Omega)$ of S by

$$(S^*\mu, \eta) = \langle \mu, S\eta \rangle_{\mathcal{M}(\Omega)} \quad \forall \mu \in \mathcal{M}(\Omega), \eta \in L^2(\Omega).$$

Note that the control-to-state operator S has the following characterisation.

LEMMA 2.1 For $\eta \in L^2(\Omega)$, $y = S\eta$ if and only if $y \in C_0(\Omega)$ satisfies

$$\forall q \in \left(n, \frac{2n}{n-2}\right) : \quad y \in W_0^{1,q}(\Omega), \quad a(y, v) = (\eta, v) \quad \forall v \in W_0^{1,q'}(\Omega). \quad (2.4)$$

Here (η, v) makes sense since $q \in (n, \frac{2n}{n-2})$ if and only if $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, and Remark 2.1 gives that for such q' we have $v \in W_0^{1,q'}(\Omega) \subset L^2(\Omega)$.

Proof. Suppose $y = S\eta$ (i.e. it solves (2.2)) and take $q \in (n, \frac{2n}{n-2})$. Since $y \in H^2(\Omega)$ we can integrate $a(y, v)$ by parts against $v \in C_c^\infty(\Omega)$ to get

$$a(y, v) = (Ay, v) \quad \forall v \in C_c^\infty(\Omega). \quad (2.5)$$

Then using (2.2) we get

$$(\eta, v) = (Ay, v) \quad \forall v \in C_c^\infty(\Omega), \quad (2.6)$$

which implies that $Ay = \eta$ a.e. in Ω . Moreover, it follows from (2.5) and the density of $C_c^\infty(\Omega)$ in $W_0^{1,q'}(\Omega)$ that $a(y, v) = (Ay, v)$ for all $v \in W_0^{1,q'}(\Omega)$. Combining this fact, $Ay = \eta$ a.e. in Ω and $v \in W_0^{1,q'}(\Omega) \subset L^2(\Omega)$ gives $a(y, v) = (\eta, v)$ for all $v \in W_0^{1,q'}(\Omega)$. By Remark 2.1 note that $y \in H^2(\Omega) \cap H_0^1(\Omega) \subset W_0^{1,q}(\Omega)$. The above arguments hold for any $q \in (n, \frac{2n}{n-2})$, so we have proved that $y = S\eta$ implies (2.4) holds.

The reverse implication is also true. Since $H_0^1(\Omega) \subset W_0^{1,q'}(\Omega)$ for any $q \in (n, \frac{2n}{n-2})$, we can test (2.4) with any $v \in H_0^1(\Omega)$. So a solution of this must solve (2.2). This completes the proof. \square

We can use this result to prove that the adjoint operator S^* can be characterised in the following way.

LEMMA 2.2 For $\mu \in \mathcal{M}(\Omega)$, $p = S^*\mu$ if and only if $p \in L^2(\Omega)$ satisfies

$$\forall q' \in \left(\frac{2n}{n+2}, \frac{n}{n-1}\right) : \quad p \in W_0^{1,q'}(\Omega), \quad a(v, p) = \int_\Omega v d\mu \quad \forall v \in W_0^{1,q}(\Omega). \quad (2.7)$$

Moreover,

$$\|p\|_{W_0^{1,q'}(\Omega)} \leq C(q')\|\mu\|_{\mathcal{M}(\Omega)} \quad \forall q' \in \left(\frac{2n}{n+2}, \frac{n}{n-1}\right). \quad (2.8)$$

Proof. Suppose (2.7) is true. Fix some $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$ then for all $\mu \in \mathcal{M}(\Omega)$ and $\eta \in L^2(\Omega)$, testing (2.7) with $S\eta \in W_0^{1,q'}(\Omega)$ gives

$$a(S\eta, p) = \int_{\Omega} S\eta \, d\mu = \langle \mu, S\eta \rangle_{\mathcal{M}(\Omega)}.$$

By the definition of S we can test (2.4) with $p \in W_0^{1,q'}(\Omega)$ to get

$$a(S\eta, p) = (\eta, p) = (p, \eta).$$

Combining these two equalities and recalling that μ and η are arbitrary we get

$$\langle \mu, S\eta \rangle_{\mathcal{M}(\Omega)} = (\eta, p) \quad \forall \mu \in \mathcal{M}(\Omega), \eta \in L^2(\Omega).$$

Comparing this to the definition of the adjoint we see $p = S^*\mu$. Since q' was arbitrary we have shown (2.7) implies $p = S^*\mu$. The uniqueness of the adjoint operator proves the reverse implication.

For the proof of the stability estimate (2.8) see Theorem 2 in Casas (1985). \square

REMARK 2.2 We have assumed that the state equation is an elliptic PDE with Dirichlet boundary conditions. The theory in this paper can be adapted to elliptic PDEs with suitable Neumann boundary conditions, provided that $a(\cdot, \cdot)$ is still coercive. This is because the same regularity results hold for them and the same error estimates hold for their finite element approximations.

3. Problem formulation

We are now in a position to formulate the optimal control problem precisely:

$$\begin{aligned} \min \quad & J(y, \eta) := \frac{1}{2} \sum_{\omega \in I} (y(\omega) - g_{\omega})^2 + \frac{\nu}{2} \|\eta\|_{L^2(\Omega)}^2 \\ \text{over} \quad & C_0(\Omega) \times L^2(\Omega) \\ \text{s.t.} \quad & y = S\eta \text{ (i.e. (2.2) holds)} \\ \text{and} \quad & \eta \in U_{ad} := \{\eta \in L^2(\Omega) : a \leq \eta \leq b \text{ a.e. in } \Omega\}. \end{aligned} \tag{3.1}$$

Recall that $I \subset \Omega$ is a finite set of points, $\{g_{\omega}\}_{\omega \in I}$ are prescribed values at these points, and $\nu > 0$. We will prove results for the case that a and b are constant real numbers with $a < b$, and also the case of no control constraints (i.e. $b = -a = \infty$).

We can use the control-to-state operator S to define the reduced objective functional $\hat{J}(\eta) = J(S\eta, \eta)$. Then it is straightforward to show that (3.1) is equivalent to the optimisation problem:

$$\begin{aligned} \min \quad & \hat{J}(\eta) = \frac{1}{2} \sum_{\omega \in I} (S\eta(\omega) - g_{\omega})^2 + \frac{\nu}{2} \|\eta\|_{L^2(\Omega)}^2 \\ \text{over} \quad & \eta \in U_{ad}. \end{aligned} \tag{3.2}$$

This equivalence is in the sense that $u \in U_{ad}$ solves (3.2) if and only if (Su, u) solves (3.1). It is simpler to work with the optimisation problem (3.2) for proving existence and uniqueness of a solution and deriving an optimality condition.

THEOREM 3.1 Problem (3.2) has a unique solution $u \in U_{ad}$, hence (3.1) has a unique solution (Su, u) .

Proof. This result follows using the same argument as is used for proving existence and uniqueness of solutions to standard optimal control problems. See e.g. Theorem 2.14 in Tröltzsch (2010) for the details. \square

THEOREM 3.2 $u \in U_{ad}$ is a solution of (3.2) if and only if there exists a $p \in L^2(\Omega)$ such that for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, $p \in W_0^{1,q'}(\Omega)$ and

$$u \in U_{ad}, \quad (p + v u, v - u) \geq 0 \quad \forall v \in U_{ad}, \quad (3.3a)$$

$$a(v, p) = \sum_{\omega \in I} (S u(\omega) - g_\omega) v(\omega) \quad \forall v \in W_0^{1,q}(\Omega). \quad (3.3b)$$

Proof. $\hat{f} : L^2(\Omega) \rightarrow \mathbb{R}$ has a Gâteaux derivative $J' : L^2(\Omega) \rightarrow L^2(\Omega)^*$. It is also (strictly) convex, and U_{ad} is a nonempty and convex subset of a real Banach space. So by a standard result (see e.g. Lemma 2.21 in Tröltzsch (2010)) $u \in L^2(\Omega)$ is a solution of (3.2) iff

$$u \in U_{ad}, \quad \langle \hat{f}'(u), v - u \rangle_{L^2(\Omega)^*, L^2(\Omega)} \geq 0 \quad \forall v \in U_{ad}. \quad (3.4)$$

For notational convenience define a function $g_d \in C^\infty(\bar{\Omega})$ such that $g_d(\omega) = g_\omega$ for all $\omega \in I$; such a function could be constructed using a mollifier. Let $\mu := \sum_{\omega \in I} \delta_\omega$, where δ_ω are Dirac measures centred at points ω , so $\mu \in \mathcal{M}(\Omega)$. Since $(Su - g_d)^2 \in C(\bar{\Omega})$ we can rewrite \hat{f} as

$$\hat{f}(u) = \frac{1}{2} \int_{\Omega} (Su - g_d)^2 d\mu + \frac{v}{2} \|u\|_{L^2(\Omega)}^2$$

and use the ideas from Casas (1986). As a result our proof applies to objective functionals of this form with arbitrary $\mu \in \mathcal{M}(\Omega)$.

Calculating \hat{f}' we find that (3.4) becomes

$$\int_{\Omega} (Su - g_d) S(v - u) d\mu + v(u, v - u) \geq 0 \quad \forall v \in U_{ad}.$$

We now show that the first term on the left hand side can be written in the form $\int_{\Omega} p(v - u) dx$, where p satisfies (3.3b).

For $u \in L^2(\Omega)$, $Su - g_d \in C(\bar{\Omega})$ and so it is measurable with respect to μ . So we can define a real Borel measure $\lambda_u : \mathcal{B} \rightarrow \mathbb{R}$ (where \mathcal{B} denotes the Borel σ -algebra of Ω) by

$$\lambda_u(A) := \int_A (Su - g_d) d\mu \quad \forall A \in \mathcal{B}. \quad (3.5)$$

Since μ is regular, we can check that λ_u is also regular. So λ_u is a real regular Borel measure (i.e. it belongs to $\mathcal{M}(\Omega)$) and Theorem 1.29 in Rudin (1987) says that for $z \in C_0(\Omega)$,

$$\int_{\Omega} (Su - g_d) z d\mu = \int_{\Omega} z d\lambda_u. \quad (3.6)$$

In particular, we can take $z := S(v - u)$ to get

$$\int_{\Omega} (Su - g_d) S(v - u) d\mu = \int_{\Omega} S(v - u) d\lambda_u = (S^* \lambda_u, v - u).$$

Let $p := S^* \lambda_u \in L^2(\Omega)$ then by (2.7), for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, $p \in W_0^{1,q'}(\Omega)$ and

$$a(v, p) = \int_{\Omega} v d\lambda_u \quad \forall v \in W_0^{1,q}(\Omega).$$

To finish, note that

$$\int_{\Omega} v d\lambda_u = \int_{\Omega} (Su - g_d) v d\mu = \sum_{\omega \in I} (Su(\omega) - g_{\omega}) v(\omega).$$

□

COROLLARY 3.1 If $u \in U_{ad}$ is a solution of (3.2) then it has the additional regularity that $u \in W^{1,q'}(\Omega)$ for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$.

Proof. Observe that (3.3a) is equivalent to

$$u(x) = \mathbb{P}_{[a,b]} \left(-\frac{1}{v} p(x) \right) \quad \text{for a.e. } x \in \Omega, \quad (3.7)$$

where $\mathbb{P}_{[a,b]}(v) := v + \max(0, a - v) - \max(0, v - b)$. If $v, w \in W^{1,q'}(\Omega)$ then $\max(v, w) \in W^{1,q'}(\Omega)$ (see e.g. Morrey Jr. (1966)). So since $p \in W^{1,q'}(\Omega)$ for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, we also get this additional regularity for u . □

3.1 Link to pointwise state constraints

We now discuss a link between the problem we consider in this paper, which penalises deviation of the state from certain values at points, and an optimal control problem with a finite number of point state constraints i.e. a problem that forces the state to take certain values at points.

Consider the following problem, which is a generalisation of (3.1) in the case of no control constraints ($b = -a = \infty$):

$$\begin{aligned} \min \quad & J_V^{\theta}(y, \eta) := \frac{1}{2} \sum_{\omega \in I} (y(\omega) - g_{\omega})^2 + v \left(\frac{1}{2} \theta \|y - g_d\|_{L^2(\Omega)}^2 + \frac{1}{2} \|\eta\|_{L^2(\Omega)}^2 \right) \\ \text{over} \quad & C_0(\Omega) \times L^2(\Omega) \\ \text{s.t.} \quad & (2.2) \text{ holds.} \end{aligned} \quad (3.8)$$

The modification is the addition of an optional $L^2(\Omega)$ fidelity term containing $g_d \in L^2(\Omega)$, which is weighted by $\theta \geq 0$. This allows us to relate (3.8) to a problem with point state constraints that is considered in the literature: In the limit $v \rightarrow 0$ we get convergence of solutions of (3.8) to the solution of the following problem, which can be found, for example, in Leykekhman *et al.* (2013):

$$\begin{aligned} \min \quad & J^{\theta}(y, \eta) := \frac{1}{2} \theta \|y - g_d\|_{L^2(\Omega)}^2 + \frac{1}{2} \|\eta\|_{L^2(\Omega)}^2 \\ \text{over} \quad & H_0^1(\Omega) \times L^2(\Omega) \\ \text{s.t.} \quad & (2.2) \text{ holds and } y(\omega) = g_{\omega} \text{ for } \omega \in I. \end{aligned} \quad (3.9)$$

THEOREM 3.3 Let (Su_v, u_v) solve (3.8) for $v > 0$ and $(S\bar{u}, \bar{u})$ solve (3.9). Then as $v \rightarrow 0$,

$$\begin{aligned} Su_v &\rightharpoonup S\bar{u} \quad \text{in } H_0^1(\Omega) \\ u_v &\rightharpoonup \bar{u} \quad \text{in } L^2(\Omega). \end{aligned}$$

Proof. First note that there exists a function $\hat{u} \in L^2(\Omega)$ such that $S\hat{u}(\omega) = g_\omega$ for all $\omega \in I$ (see Lemma 1 in Leykekhman *et al.* (2013)), so $J_V^\theta(Su_V, u_V) \leq v J^\theta(S\hat{u}, \hat{u})$. For all $v > 0$, $(S\hat{u}, \hat{u})$ is feasible for (3.8) so

$$\frac{v}{2} \|u_V\|_{L^2(\Omega)}^2 \leq J_V^\theta(Su_V, u_V) \leq v J^\theta(S\hat{u}, \hat{u}) \leq Cv \quad (3.10)$$

with C independent of v . So u_V is uniformly bounded with respect to v in $L^2(\Omega)$, which means for every sequence $v_k \rightarrow 0$ there exists a sequence $u_{v_k} \rightarrow \tilde{u}$ in $L^2(\Omega)$. Moreover (3.10) and the stability result

$$\|Su_{v_k}\|_{H_0^1(\Omega)} \leq C \|u_{v_k}\|_{L^2(\Omega)}$$

with C independent of u_{v_k} allows us to find a further subsequence, which we also denote by $\{v_k\}$, such that $Su_{v_k} \rightarrow \tilde{y}$ in $H_0^1(\Omega)$. Then taking the limit in (2.2) we see that $\tilde{y} = S\tilde{u}$. To complete the proof we need to show that $\tilde{u} = \bar{u}$, which we do by showing that $(S\tilde{u}, \tilde{u})$ is feasible for (3.9) and that $J^\theta(S\tilde{u}, \tilde{u}) \leq J^\theta(S\bar{u}, \bar{u})$.

Note that the same reasoning as for (3.10) gives $\frac{1}{v} \sum_{\omega \in I} (Su_V(\omega) - g_\omega)^2 \leq C$ independently of v . Therefore we must have $Su_V(\omega) \rightarrow g_\omega$. So $S\tilde{u}(\omega) = g_\omega$ for $\omega \in I$ and $(S\tilde{u}, \tilde{u})$ is feasible for (3.9).

The weak lower semicontinuity of J^θ and $J^\theta(Su_{v_k}, u_{v_k}) \leq \frac{J_{v_k}^\theta(Su_{v_k}, u_{v_k})}{v_k}$ implies

$$J^\theta(S\tilde{u}, \tilde{u}) \leq \liminf_{k \rightarrow \infty} J^\theta(Su_{v_k}, u_{v_k}) \leq \liminf_{k \rightarrow \infty} \frac{J_{v_k}^\theta(Su_{v_k}, u_{v_k})}{v_k}.$$

Also the optimality of (Su_{v_k}, u_{v_k}) for (3.8) and $\frac{J_{v_k}^\theta(S\bar{u}, \bar{u})}{v_k} = J^\theta(S\bar{u}, \bar{u})$ implies

$$\liminf_{k \rightarrow \infty} \frac{J_{v_k}^\theta(Su_{v_k}, u_{v_k})}{v_k} \leq \liminf_{k \rightarrow \infty} \frac{J_{v_k}^\theta(S\bar{u}, \bar{u})}{v_k} = J^\theta(S\bar{u}, \bar{u}).$$

Combining these we get

$$J^\theta(S\tilde{u}, \tilde{u}) \leq J^\theta(S\bar{u}, \bar{u}),$$

so we have proved the result. \square

4. Discretisation

In this section we discretise the state equation using a finite element method and use this to formulate two different discrete problems. We then derive discrete optimality conditions for each problem.

We now make slightly stronger assumptions on Ω than were necessary for the problem formulation and analysis in the previous section. From now onwards assume that Ω is convex with a C^2 boundary. The assumption of convexity simplifies the presentation since then the finite element space for the state (defined shortly) is a subset of $C_0(\Omega)$. Note that if the state equation had Neumann boundary conditions (see Remark 2.2) then nonconvex domains would not cause this complication. A C^2 boundary is sufficiently smooth that for $2 \leq s < \infty$,

$$\|S\eta\|_{W^{2,s}(\Omega)} \leq C(s) \|\eta\|_{L^s(\Omega)} \quad \forall \eta \in L^s(\Omega) \quad (4.1)$$

(see e.g. Theorems 9.14 and 9.15 in Gilbarg & Trudinger (2001)).

We can take a family of polygonal approximations $\Omega_h \subset \Omega$ such that the vertices of $\partial\Omega_h$ lie on $\partial\Omega$ and $|\Omega \setminus \Omega_h| \leq Ch^2$. On each Ω_h we can construct a conforming triangulation T_h of triangles or

tetrahedra T with maximum diameter $h := \max_{T \in T_h} h(T)$, where $h(T)$ is the diameter of an element T . Additionally suppose that the family of triangulations are conforming and quasi-uniform i.e. there exists a constant C such that

$$\frac{h(T)}{\rho(T)} \leq C \quad \forall T \in T_h,$$

where $\rho(T)$ is the radius of the largest ball contained in T , and there exists a constant C such that

$$\frac{h}{h(T)} \leq C \quad \forall T \in T_h$$

(see e.g. Chapter 3 in Ciarlet (1978)). We can define the following family of discrete spaces of piecewise linear globally continuous finite elements which vanish on the boundary:

$$V_h := \{v_h \in C_0(\Omega) : v_h|_T \in P_1(T) \text{ for all } T \in T_h \text{ and } v_h|_{\Omega \setminus \Omega_h} = 0\}.$$

Here $P_1(T)$ is the set of affine functions over T . Our motivation for using this finite element space (rather than, for example, a space of piecewise constant finite elements) is that it is a subspace of $C_0(\Omega)$.

We also construct a family of triangulations T^σ of triangles or tetrahedra with maximum element diameter σ . We allow elements on the boundary to have one curved face, and assume that T^σ is conforming and shape regular (as we did for T_h). Note that the family of triangulations T^σ potentially has nothing in common with T_h . We can now define the following discrete space $U_{ad,\sigma}$ for the control:

$$\begin{aligned} U_\sigma &:= \{u_\sigma \in C(\bar{\Omega}) : u_\sigma|_T \in P_1(T) \text{ for all } T \in T^\sigma\}, \\ U_{ad,\sigma} &:= \{u_\sigma \in U_\sigma : a \leq u_\sigma \leq b\}. \end{aligned}$$

This is a space of piecewise linear globally continuous finite elements (as was V_h) with $U_{ad,\sigma} \subset U_{ad}$, however we do not require the functions to vanish at the boundary. Recall from Corollary 3.1 that $u \in W^{1,q'}(\Omega)$ for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, and piecewise constant finite elements approximate such functions almost as well as piecewise linear finite elements. However we define $U_{ad,\sigma}$ to use piecewise linear finite elements as we want to allow taking the same discrete space for the control and state. This can simplify implementations.

For U_σ the following approximation property holds: There exists an interpolation operator $\Pi_\sigma : W^{l,p}(\Omega) \rightarrow U_\sigma$ ($l = 1, 2; 1 \leq p < \infty$) such that

$$\|v - \Pi_\sigma v\|_{W^{m,p}(\Omega)} \leq C \sigma^{1-m} \|v\|_{W^{1,p}(\Omega)}, \quad m = 0, 1. \quad (4.2)$$

Such an interpolation operator can be defined as in Scott & Zhang (1990). It also has the property that $v \in U_{ad}$ implies $\Pi_\sigma v \in U_{ad}$.

We now introduce discrete approximations of S and S^* . Define $S_h : L^2(\Omega) \rightarrow C_0(\Omega)$ by $S_h \eta = y_h$, where y_h satisfies

$$y_h \in V_h, \quad a(y_h, v_h) = (\eta, v_h) \quad \forall v_h \in V_h. \quad (4.3)$$

It is a standard result that this problem has a unique solution. We now prove some estimates for S_h that will be useful for the numerical analysis.

LEMMA 4.1 For $\eta \in L^s(\Omega)$ and $2 \leq s < \infty$,

$$\|S\eta - S_h \eta\|_\infty \leq C(s) h^{2-\frac{n}{s}} \|\eta\|_{L^s(\Omega)}, \quad n = 2, 3. \quad (4.4)$$

Proof. First we will recall some results from the literature that hold under the assumptions we make in this paper. By (34) in Leykekhman *et al.* (2013) we have that

$$\|Sv - S_h v\|_{L^s(\Omega)} \leq C(s)h^2 \|Sv\|_{W^{2,s}(\Omega)} \quad \forall v \in L^s(\Omega).$$

This was originally proved for $n = 2$ on p438 in Rannacher & Scott (1982). Applying an inverse inequality on each element of the triangulation gives that

$$\|v_h\|_{L^\infty(\Omega_h)} \leq C(s)h^{-\frac{n}{s}} \|v_h\|_{L^s(\Omega)} \quad \forall v_h \in V_h \quad (4.5)$$

(see e.g. Ciarlet (1978)). Similarly, for the piecewise linear interpolation operator $I_h : C_0(\Omega) \rightarrow V_h$ and $r \in [1, \infty]$ we have

$$\|v - I_h v\|_{L^r(\Omega_h)} \leq C(s)h^{2+\frac{1}{r}-\frac{1}{s}} \|v\|_{W^{2,s}(\Omega_h)} \quad \forall v \in W^{2,s}(\Omega).$$

(see e.g. Theorem 3.1.5 in Ciarlet (1978)).

Combining these results we get that

$$\begin{aligned} \|S\eta - S_h \eta\|_{L^\infty(\Omega_h)} &\leq \|S\eta - I_h S\eta\|_{L^\infty(\Omega_h)} + \|I_h S\eta - S_h \eta\|_{L^\infty(\Omega_h)}, \\ &\leq C(s)(h^{2-\frac{n}{s}} \|S\eta\|_{W^{2,s}(\Omega)} + h^{-\frac{n}{s}} \|I_h S\eta - S_h \eta\|_{L^s(\Omega_h)}) \\ &\leq C(s)h^{-\frac{n}{s}} (h^2 \|S\eta\|_{W^{2,s}(\Omega)} + \|I_h S\eta - S\eta\|_{L^s(\Omega_h)} + \|S\eta - S_h \eta\|_{L^s(\Omega_h)}) \\ &\leq C(s)h^{2-\frac{n}{s}} (\|S\eta\|_{W^{2,s}(\Omega)} + \|\eta\|_{L^s(\Omega)}) \\ &\leq C(s)h^{2-\frac{n}{s}} \|\eta\|_{L^s(\Omega)}. \end{aligned}$$

We now need to prove a supremum norm error estimate for the skin $\Omega \setminus \Omega_h$. By Theorem 4.12 Part II in Adams & Fournier (2003):

- If $s \geq n$ then $W^{2,s}(\Omega) \hookrightarrow C^{0,\lambda}(\bar{\Omega})$ for $0 < \lambda < 1$.
- If $\frac{n}{2} < s < n$ then $W^{2,s}(\Omega) \hookrightarrow C^{0,\lambda}(\bar{\Omega})$ for $0 < \lambda \leq 2 - \frac{n}{s}$.

Let

$$\bar{\lambda}(s) := \begin{cases} 1 - \frac{n}{2s} & s \geq n, \\ 2 - \frac{n}{s} & \frac{n}{2} \leq s < n, \end{cases}$$

and observe that for $x_1 \in \Omega \setminus \Omega_h$ we have

$$\inf_{x_2 \in \partial\Omega} |S\eta(x_1) - S\eta(x_2)| \leq C(s) \inf_{x_2 \in \partial\Omega} |x_1 - x_2|^{\bar{\lambda}(s)}.$$

From the smoothness of the domain it follows that

$$\inf_{x_2 \in \partial\Omega} |x_1 - x_2| \leq Ch^2 \quad \forall x_1 \in \Omega \setminus \Omega_h$$

and for $2 \leq s < \infty$ we have $h^{2\bar{\lambda}(s)} \leq Ch^{2-\frac{n}{s}}$ for sufficiently small h . Using this and $S\eta|_{\partial\Omega} = 0$ we get

$$|S\eta(x_1)| \leq C(s)h^{2-\frac{n}{s}} \quad \forall x_1 \in \Omega \setminus \Omega_h.$$

Hence

$$\|S\eta - S_h \eta\|_\infty \leq \max(\|S\eta - S_h \eta\|_{L^\infty(\Omega_h)}, \|S\eta - S_h \eta\|_{L^\infty(\Omega \setminus \Omega_h)}) \leq C(s)h^{2-\frac{n}{s}} \|\eta\|_{L^s(\Omega)}.$$

□

COROLLARY 4.1 For $\eta \in L^2(\Omega)$,

$$\|S\eta - S_h\eta\|_\infty \leq Ch^{2-\frac{n}{2}} \|\eta\|_{L^2(\Omega)}, \quad n = 2, 3. \quad (4.6)$$

For $\eta \in W^{1,q'}(\Omega)$ with $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\|S\eta - S_h\eta\|_\infty \leq C(q')h^{3-\frac{n}{q'}} \|\eta\|_{W^{1,q'}(\Omega)} \quad n = 2, 3. \quad (4.7)$$

Proof. The first estimate follows by taking $s = 2$ in Lemma 4.1. The other estimate follow by combining the lemma with Sobolev embedding results. In particular, if $\eta \in W^{1,q'}(\Omega)$ with $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$ then $W^{1,q'}(\Omega) \hookrightarrow L^s(\Omega)$ with $s = \frac{nq'}{n-q'} \geq 2$. So

$$C(s)h^{2-\frac{n}{s}} \|\eta\|_{L^s(\Omega)} \leq C(q')h^{3-\frac{n}{q'}} \|\eta\|_{W^{1,q'}(\Omega)},$$

which proves the second estimate. Note that this estimate is proved in a similar way in Theorem 3 in Leykekhman *et al.* (2013). \square

We will use (4.6) in Section 5.1 and (4.7) in Section 5.2 to prove $L^2(\Omega)$ error estimates for the point optimal control problem.

Since S_h is continuous (which follows from (4.6)) and linear it has an adjoint operator $S_h^* : \mathcal{M}(\Omega) \rightarrow L^2(\Omega)$. Note that the same calculation as in Lemma 2.2 gives that $p_h = S_h^*\mu$ if and only if p_h satisfies

$$p_h \in V_h, \quad a(v_h, p_h) = \int_\Omega v_h d\mu \quad \forall v_h \in V_h. \quad (4.8)$$

We have the following error estimate for S_h^* , which we will use in Section 5.1: For $\mu \in \mathcal{M}(\Omega)$,

$$\|S^*\mu - S_h^*\mu\|_{L^2(\Omega)} \leq Ch^{2-\frac{n}{2}} \|\mu\|_{\mathcal{M}(\Omega)}, \quad (4.9)$$

with C independent of μ and h . This follows by noting that for any $v \in L^2(\Omega)$,

$$(S^*\mu - S_h^*\mu, v) = \langle \mu, Sv - S_hv \rangle_{\mathcal{M}(\Omega)} \leq \|\mu\|_{\mathcal{M}(\Omega)} \|Sv - S_hv\|_\infty.$$

Then using (4.6) gives the result. The estimate (4.9) was originally proved for convex polygonal domains in Theorem 3 in Casas (1985), and related theory is developed in Scott (1973).

REMARK 4.1 The estimates in Lemma 4.1 and Corollary 4.1 still hold if S and S_h are appropriately defined control-to-state operators corresponding to an elliptic PDE with Neumann boundary conditions.

4.1 Discrete problems

We are now ready to introduce the two discrete problems that we consider in our numerical analysis.

Define the discrete reduced objective functional $\hat{J}_h : L^2(\Omega) \rightarrow \mathbb{R}$ by

$$\hat{J}_h(\eta) = J(S_h\eta, \eta) = \frac{1}{2} \sum_{\omega \in I} (S_h\eta(\omega) - g_\omega)^2 + \frac{\nu}{2} \|\eta\|_{L^2(\Omega)}^2.$$

Then the first discrete problem we consider is (M1_h):

$$\min \hat{J}_h(\eta_\sigma) \text{ over } \eta_\sigma \in U_{ad,\sigma}. \quad (4.10)$$

PROPOSITION 4.1 There is a unique solution $u_{\sigma,h} \in U_{ad,\sigma}$ to $(M1_h)$ (see (4.10)). Moreover, $u_{\sigma,h} \in U_{ad,\sigma}$ is a solution of $(M1_h)$ if and only if there exists $p_h \in V_h$ such that

$$u_{\sigma,h} \in U_{ad,\sigma}, \quad (p_h + v u_{\sigma,h}, v_{\sigma} - u_{\sigma,h}) \geq 0 \quad \forall v_{\sigma} \in U_{ad,\sigma} \quad (4.11a)$$

$$a(v_h, p_h) = \sum_{\omega \in I} (S_h u_{\sigma,h}(\omega) - g_{\omega}) v_h(\omega) \quad \forall v_h \in V_h. \quad (4.11b)$$

Proof. The proof follows from the same considerations as in Theorems 3.1 and 3.2. Note that $p_h = S_h^* \lambda_{h,u_{\sigma,h}}$ where for $\eta \in L^2(\Omega)$ we define $\lambda_{h,\eta} \in \mathcal{M}(\Omega)$ by

$$\lambda_{h,\eta}(A) = \int_A (S_h \eta - g_d) d\mu \quad \forall A \in \mathcal{B} \quad (4.12)$$

with $\mu = \sum_{\omega \in I} \delta_{\omega}$. \square

We refer to $(M1_h)$ as the explicitly discretised problem as we make the control belong to a space of discrete functions.

Alternatively we could use the variational discretisation concept from Hinze (2005) and leave the control in the infinite dimensional space U_{ad} . This leads to the potentially different (see Remark 4.2) discrete problem $(M2_h)$:

$$\min \hat{J}_h(\eta) \text{ over } \eta \in U_{ad}. \quad (4.13)$$

PROPOSITION 4.2 There is a unique solution $u_h \in U_{ad}$ to $(M2_h)$ (see (4.13)). Moreover, $u_h \in L^2(\Omega)$ is a solution of $(M2_h)$ if and only if there exists $p_h \in V_h$ such that

$$u_h \in U_{ad}, \quad (p_h + v u_h, v - u_h) \geq 0 \quad \forall v \in U_{ad} \quad (4.14a)$$

$$a(v_h, p_h) = \sum_{\omega \in I} (S_h u_h(\omega) - g_{\omega}) v_h(\omega) \quad \forall v_h \in V_h. \quad (4.14b)$$

Proof. The proof also follows from the same considerations as in Theorems 3.1 and 3.2. \square

A priori we only know that u_h belongs to U_{ad} . However observe that (4.14a) can be expressed using the pointwise projection operator $\mathbb{P}_{[a,b]}$ from (3.7) as

$$u_h = \mathbb{P}_{[a,b]} \left(-\frac{1}{v} p_h \right).$$

So (4.14a) has a simpler form than (4.11a), which is an $L^2(\Omega)$ projection onto a discrete space. This means u_h inherits a piecewise linear structure from $p_h \in V_h$, but observe that u_h does not necessarily belong to V_h due to the control constraints. We refer to this as an implicit discretisation; we are not requiring u_h to be a piecewise linear function, but it gains this property indirectly through the discretisation of the state. Even though u_h does not necessarily belong to V_h , this problem can be solved computationally. We will elaborate on this in Section 6.1.

REMARK 4.2 The motivation for the implicitly discretised problem $(M2_h)$ is that it allows a better approximation of the set where the control constraints are active (indicated in Figure 1), likely leading to a smaller error. For a more thorough explanation see Hinze (2005).

REMARK 4.3 Note that if there are no active control constraints (e.g. if $b = -a = \infty$) and $V_h \subset U_{\sigma}$, then $(M1_h)$ and $(M2_h)$ are equivalent. In order for $V_h \subset U_{\sigma}$ we need “ $T^{\sigma} \subset T_h$ ”. By this we mean that each element of T^{σ} is contained in either a single element of T_h or the skin $\Omega \setminus \Omega_h$.

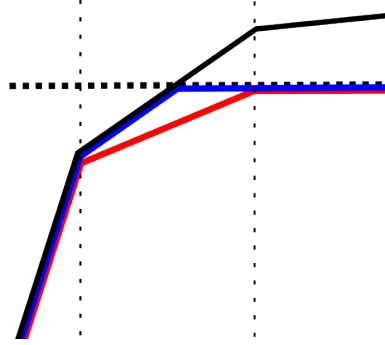


FIG. 1. An illustration in 1D of how u_h is determined by p_h (black line) when the discrete space for the control and state are the same and $v = 1$. The horizontal dashed line is the value of b and the vertical dashed lines marks the boundary between elements. The blue line is u_h calculated from p_h using (4.11a) and the red line is using (4.14a). Assuming the p_h are similar and good approximations of p for both $(M1_h)$ and $(M2_h)$, this suggests that $(M2_h)$ will give a better approximation of u .

5. Numerical analysis

We now prove $L^2(\Omega)$ error estimates between the solution of the continuous problem (3.2) and the two discrete problems $(M1_h)$ and $(M2_h)$ (see (4.10) and (4.13)). We use two different approaches for this numerical analysis. Approach 1 in the next section allows us to prove error estimates for the two discrete problems in most (but not all) the cases we would like. Approach 2 in Section 5.2 only reproduces some of these error estimates, however it is simpler.

5.1 Approach 1

This error analysis is based on Casas & Tröltzsch (2003), where an a priori $L^2(\Omega)$ error estimate is proved for the standard optimal control problem which has an $L^2(\Omega)$ fidelity term. The approach allows us to prove $L^2(\Omega)$ error estimates for both $(M1_h)$ and $(M2_h)$. The only estimates it does not give are ones for $(M1_h)$ when $n = 3$ (but we are not able to prove these using Approach 2 either). In particular we will get the following results.

THEOREM 5.1 Assume $n = 2$. Let u solve (3.2) and $u_{\sigma,h}$ solve $(M1_h)$ (see (4.10)). Then

$$\|u - u_{\sigma,h}\|_{L^2(\Omega)} \leq C(\sqrt{\sigma} + h)$$

with C independent of σ and h .

THEOREM 5.2 Assume $n = 2$ or 3 . Let u solve (3.2) and u_h solve $(M2_h)$ (see (4.13)). Then

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{2-\frac{n}{2}}$$

with C independent of σ and h .

COROLLARY 5.1 Assume $n = 2$ or 3 , there are no active control constraints (e.g. $b = -a = \infty$), and that $V_h \subset U_\sigma$. Let u solve (3.2) and $u_{\sigma,h}$ solve $(M1_h)$ (see (4.10)). Then

$$\|u - u_{\sigma,h}\|_{L^2(\Omega)} \leq Ch^{2-\frac{n}{2}}$$

with C independent of σ and h .

Proof. This result follows from the equivalence between (M1_h) and (M2_h) that is highlighted in Remark 4.3. \square

Note that these results suggest (M2_h) is the preferred discretisation. In particular, we can only prove an error estimate in the case of $n = 3$ with control constraints for (M2_h). Also the error estimate in the case of $n = 2$ with control constraints is better for (M2_h).

The idea of the approach is to consider the error caused by the discretisation of the control and state separately, then combine them. This approach only needs the weak supremum norm error estimate for the state equation (where as a stronger one is needed for Approach 2 in Section 5.2), but it does require an error estimate for the adjoint of the control-to-state operator. An advantage of this approach is that it can give insight into the best choice of triangulations for the control and state, which are not necessarily the same.

To begin we split the error as

$$\|u - u_{\sigma,h}\|_{L^2(\Omega)} \leq \|u - u_{\sigma}\|_{L^2(\Omega)} + \|u_{\sigma} - u_{\sigma,h}\|_{L^2(\Omega)}, \quad (5.1)$$

where u_{σ} solves the semi discrete control problem

$$\min \hat{J}(\eta_{\sigma}) \text{ over } \eta_{\sigma} \in U_{ad,\sigma}. \quad (5.2)$$

PROPOSITION 5.3 There is a unique solution $u_{\sigma} \in U_{ad,\sigma}$ to (5.2). Moreover, $u_{\sigma} \in U_{ad,\sigma}$ is a solution of (5.2) if and only if there exists a $p_{\sigma} \in L^2(\Omega)$ such that for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, $p_{\sigma} \in W_0^{1,q'}(\Omega)$ and

$$u \in U_{ad,\sigma} \quad (p_{\sigma} + v u_{\sigma}, v_{\sigma} - u_{\sigma}) \geq 0 \quad \forall v_{\sigma} \in U_{ad,\sigma} \quad (5.3a)$$

$$a(v, p_{\sigma}) = \sum_{\omega \in I} (S u_{\sigma}(\omega) - g_{\omega}) v(\omega) \quad \forall v \in W_0^{1,q}(\Omega). \quad (5.3b)$$

Note that p_{σ} is not a discrete function. The subscript σ is to denote association with the discrete control u_{σ} .

Proof. $U_{ad,\sigma}$ is still a closed convex subset of $L^2(\Omega)$ so the proofs in Theorems 3.1 and 3.2 apply. Note that $p_{\sigma} = S^* \lambda_{u_{\sigma}}$ where $\lambda_{u_{\sigma}} \in \mathcal{M}(\Omega)$ is defined analogously to (3.5) by

$$\lambda_{u_{\sigma}}(A) := \int_A (S u_{\sigma} - g_d) d\mu \quad \forall A \in \mathcal{B}. \quad (5.4)$$

\square

Whereas (4.13) minimises the discrete reduced objective functional over the continuous space, this problem minimises the continuous reduced objective functional over the discrete space. So the solution of (5.2) is discrete, but the corresponding state is continuous, and this problem cannot be solved computationally.

The first term on the right hand side of (5.1) can be thought of as the error from the discretisation of the control, as we are comparing the minimiser of the continuous objective functional over continuous and discrete controls. Similarly the second term on the right hand side of (5.1) can be thought of as the error from the discretisation of the state, as we compare the minimiser of the continuous and discrete objective functionals, both over discrete controls. To prove Theorem 5.1 it is sufficient to prove an error estimate for each term separately, which we do in Lemmas 5.2 and 5.3. Note that we have additional assumptions in Theorem 5.1 because we need these in order to prove Lemma 5.2. But first we will prove some a priori estimates for the solution of (5.2).

LEMMA 5.1 Let u_σ solve (5.2) and p_σ satisfy the optimality system (5.3). For all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\|u_\sigma\|_{L^2(\Omega)} + \|Su_\sigma\|_{L^2(\Omega)} + \|p_\sigma\|_{W_0^{1,q'}(\Omega)} \leq C(q') \quad (5.5)$$

with C independent of σ . Moreover, when $n = 2$ there exists some $q > n$ such that

$$\|u_\sigma\|_{L^q(\Omega)} + \|p_\sigma\|_{L^q(\Omega)} \leq C \quad (5.6)$$

with C independent of σ .

Proof. Using (2.8), (5.4) and (2.3), for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$ we have

$$\begin{aligned} \|p_\sigma\|_{W_0^{1,q'}(\Omega)} &\leq C(q') \|\lambda_{u_\sigma}\|_{\mathcal{M}(\Omega)} \\ &= C(q') \sum_{\omega \in I} |Su_\sigma(\omega) - g_d| \\ &\leq C(q') (\|Su_\sigma\|_\infty + \max_{\omega \in I} |g_\omega|) \\ &\leq C(q') (\|u_\sigma\|_{L^2(\Omega)} + 1). \end{aligned} \quad (5.7)$$

Combining this with (2.3) again we get

$$\|u_\sigma\|_{L^2(\Omega)} + \|Su_\sigma\|_{L^2(\Omega)} + \|p_\sigma\|_{W_0^{1,q'}(\Omega)} \leq C(q') (\|u_\sigma\|_{L^2(\Omega)} + 1). \quad (5.8)$$

If $a, b \in \mathbb{R}$ then

$$\|u_\sigma\|_{L^2(\Omega)} \leq |\Omega|^{\frac{1}{2}} \max(|a|, |b|).$$

If $b = -a = \infty$ then $0 \in U_{ad,\sigma}$, so $\hat{f}(u_\sigma) \leq \hat{f}(0)$. Since $S0 = 0$, this means

$$\frac{v}{2} \|u_\sigma\|_{L^2(\Omega)}^2 \leq \frac{1}{2} \sum_{\omega \in I} g_\omega^2. \quad (5.9)$$

So regardless of the assumptions on a and b , we have $\|u_\sigma\|_{L^2(\Omega)} \leq C$. Combining this with (5.8) gives the first bound (5.5).

For the second bound we assume $n = 2$. If $a, b \in \mathbb{R}$ then we have

$$\|u_\sigma\|_{L^q(\Omega)} \leq |\Omega|^{\frac{1}{q}} \max(|a|, |b|).$$

If $b = -a = \infty$ we can use the $L^q(\Omega)$ stability of the $L^2(\Omega)$ projection (see e.g. Crouzeix & Thomée (1987)) to get $\|u_\sigma\|_{L^q(\Omega)} \leq \frac{1}{v} \|p_\sigma\|_{L^q(\Omega)}$. So for all $q > 2$,

$$\|u_\sigma\|_{L^q(\Omega)} + \|p_\sigma\|_{L^q(\Omega)} \leq C(\|p_\sigma\|_{L^q(\Omega)} + 1).$$

We now need some $q > 2$ such that $\|p_\sigma\|_{L^q(\Omega)} \leq C$ independently of σ . By Sobolev embedding results, if $s > \frac{n}{2}$ then $W^{1,s}(\Omega) \hookrightarrow L^t(\Omega)$ for some $t > n$. In particular for $n = 2$ we can take $s = \frac{4}{3} > \frac{n}{2} = 1$, since $p_\sigma \in W_0^{1,s}(\Omega)$ for $s \in (\frac{2n}{n+2}, \frac{n}{n-1}) = (1, 2)$. Then for some $q > 2$,

$$\|p_\sigma\|_{L^q(\Omega)} \leq C \|p_\sigma\|_{W_0^{1,\frac{4}{3}}(\Omega)} \leq C,$$

where we have used (5.5) for the final inequality. Note that for $n = 3$ we would require $s > \frac{n}{2} = \frac{3}{2}$, but for such an s we do not have $p_\sigma \in W_0^{1,s}(\Omega)$, which only holds when $s \in (\frac{2n}{n+2}, \frac{n}{n-1}) = (\frac{5}{4}, \frac{3}{2})$. \square

LEMMA 5.2 (Error from discretisation of the control) Assume $n = 2$. Let u and u_σ be solutions of (3.2) and (5.2) respectively. Then

$$\|u - u_\sigma\|_{L^2(\Omega)} \leq C\sqrt{\sigma}$$

with C independent of σ (and h).

Proof. Test with $v = u_\sigma$ in (3.3a) to get

$$(p + vu, u_\sigma - u) \geq 0.$$

Test with $v_\sigma = \Pi_\sigma u$ in (5.3a) to get

$$(p_\sigma + vu_\sigma, \Pi_\sigma u - u_\sigma) = (p_\sigma + vu_\sigma, \Pi_\sigma u - u) + (p_\sigma + vu_\sigma, u - u_\sigma) \geq 0.$$

Adding these two inequalities and rearranging we get

$$v\|u - u_\sigma\|_{L^2(\Omega)}^2 + (p_\sigma - p, u_\sigma - u) \leq (p_\sigma + vu_\sigma, \Pi_\sigma u - u). \quad (5.10)$$

Recall from the proof of Theorem 3.2 that $p = S^* \lambda_u$ with λ_u defined by (3.5). Similarly during the proof of Proposition 5.3 we find that $p_\sigma = S^* \lambda_{u_\sigma}$ with λ_{u_σ} defined by (5.4). So using this and Theorem 1.29 in Rudin (1987) (see e.g. (3.6)) we get

$$\begin{aligned} (p_\sigma - p, u_\sigma - u) &= (S^* \lambda_{u_\sigma} - S^* \lambda_u, u_\sigma - u) = \langle \lambda_{u_\sigma} - \lambda_u, S(u_\sigma - u) \rangle_{\mathcal{M}(\Omega)} \\ &= \int_{\Omega} (S(u - u_\sigma))^2 d\mu \geq 0. \end{aligned}$$

This means the second term on the left hand side of (5.10) can be dropped.

We now bound the right hand side of (5.10). By Lemma 5.1, for $n = 2$ there exists some $q > n$ such that $\|u_\sigma\|_{L^q(\Omega)}$ and $\|p_\sigma\|_{L^q(\Omega)}$ are bounded independently of σ . So using Hölder's inequality with this q we get

$$\begin{aligned} (p_\sigma + vu_\sigma, \Pi_\sigma u - u) &\leq \|p_\sigma + vu_\sigma\|_{L^q(\Omega)} \|\Pi_\sigma u - u\|_{L^{q'}(\Omega)} \\ &\leq (\|p_\sigma\|_{L^q(\Omega)} + v\|u_\sigma\|_{L^q(\Omega)}) \|\Pi_\sigma u - u\|_{L^{q'}(\Omega)} \\ &\leq C \|\Pi_\sigma u - u\|_{L^{q'}(\Omega)}, \end{aligned}$$

with C independent of σ . Now (4.2) gives

$$\|\Pi_\sigma u - u\|_{L^{q'}(\Omega)} \leq C\sigma \|u\|_{W^{1,q'}(\Omega)} \leq C\sigma,$$

so we can deduce that

$$(p_\sigma + vu_\sigma, \Pi_\sigma u - u) \leq C\sigma.$$

Recall from Lemma 5.1 that $\|u_\sigma\|_{L^q(\Omega)}$ and $\|p_\sigma\|_{L^q(\Omega)}$ are not bounded independently of σ for $n = 3$, so the above proof does not work in that case. \square

LEMMA 5.3 (Error from discretisation of the state) Assume $n = 2$ or 3 . Let u_σ and $u_{\sigma,h}$ be the solutions of (5.2) and (M1_h) (see (4.10)) respectively. Then

$$\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \leq Ch^{2-\frac{n}{2}}$$

with C independent of σ and h .

Proof. Testing (4.11a) with $v_\sigma = u_\sigma$ gives

$$(p_h + v u_{\sigma,h}, u_\sigma - u_{\sigma,h}) \geq 0.$$

Testing (5.3a) with $v_h = u_{\sigma,h}$ gives

$$(p_\sigma + v u_\sigma, u_{\sigma,h} - u_\sigma) \geq 0.$$

Adding these two inequalities, using that $p_h = S_h^* \lambda_{h,u_{\sigma,h}}$ and $p_\sigma = S^* \lambda_{u_\sigma}$, and introducing $S_h^* \lambda_{h,u_\sigma}$ (see (4.12)) we get

$$\begin{aligned} v \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)}^2 &\leq (p_h - p_\sigma, u_\sigma - u_{\sigma,h}) \\ &= (S_h^* \lambda_{h,u_{\sigma,h}} - S^* \lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}) \\ &\leq (S_h^* \lambda_{h,u_{\sigma,h}} - S_h^* \lambda_{h,u_\sigma}, u_\sigma - u_{\sigma,h}) \\ &\quad + (S_h^* \lambda_{h,u_\sigma} - S^* \lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}). \end{aligned} \quad (5.11)$$

Note that

$$\begin{aligned} (S_h^* \lambda_{h,u_{\sigma,h}} - S_h^* \lambda_{h,u_\sigma}, u_\sigma - u_{\sigma,h}) &= (\lambda_{h,u_{\sigma,h}} - \lambda_{h,u_\sigma}, S_h(u_\sigma - u_{\sigma,h}))_{\mathcal{M}(\Omega)} \\ &= - \int_{\Omega} (S_h(u_{\sigma,h} - u_\sigma))^2 d\mu \leq 0. \end{aligned}$$

So the first term on the right hand side of (5.11) can be dropped. Also note that

$$(S_h^* \lambda_{h,u_\sigma} - S^* \lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}) = (S_h^* \lambda_{h,u_\sigma} - S^* \lambda_{h,u_\sigma}, u_\sigma - u_{\sigma,h}) + (S^* \lambda_{h,u_\sigma} - S^* \lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}),$$

and we can bound both terms on the right hand side of this. Using (4.9) and

$$\|\lambda_{h,u_\sigma}\|_{\mathcal{M}(\Omega)} = \sum_{\omega \in I} |S_h u_\sigma(\omega) - g_\omega| \leq C \|S_h u_\sigma\|_\infty + \max_{\omega \in I} |g_\omega| \leq C (\|u_\sigma\|_{L^2(\Omega)} + 1) \leq C, \quad (5.12)$$

we get

$$\begin{aligned} (S_h^* \lambda_{h,u_\sigma} - S^* \lambda_{h,u_\sigma}, u_\sigma - u_{\sigma,h}) &\leq C \|S_h^* \lambda_{h,u_\sigma} - S^* \lambda_{h,u_\sigma}\|_{L^2(\Omega)} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\leq C h^{2-\frac{n}{2}} \|\lambda_{h,u_\sigma}\|_{\mathcal{M}(\Omega)} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\leq C h^{2-\frac{n}{2}} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \end{aligned} \quad (5.13)$$

with C independent of σ and h . By (4.6) we have

$$\begin{aligned} (S^* \lambda_{h,u_\sigma} - S^* \lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}) &\leq C \|S^* \lambda_{h,u_\sigma} - S^* \lambda_{u_\sigma}\|_{L^2(\Omega)} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\leq C \|\lambda_{h,u_\sigma} - \lambda_{u_\sigma}\|_{\mathcal{M}(\Omega)} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &= C \left(\sum_{\omega \in I} |S_h u_\sigma(\omega) - S u_\sigma(\omega)| \right) \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\leq C \|S_h u_\sigma - S u_\sigma\|_{L^\infty(\Omega)} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\leq C h^{2-\frac{n}{2}} \|u_\sigma\|_{L^2(\Omega)} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\leq C h^{2-\frac{n}{2}} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \end{aligned} \quad (5.14)$$

with C independent of σ and h . So

$$(S^* \lambda_{h,u_\sigma} - S^* \lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}) \leq Ch^{2-\frac{n}{2}} \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)},$$

and using this in (5.11) completes the proof. \square

Combining Lemmas 5.2 and 5.3 gives Theorem 5.1. A consequence of the theorem is that in 2 dimensions by taking $\sigma = h^2$ we can get an $O(h)$ error estimate for the explicitly discretised problem $(M1_h)$. In this case the state is on a triangulation of size $O(h)$ and the control is on a triangulation of size $O(h^2)$ (i.e. the control space on a finer triangulation than the state). Even though a finer triangulation is involved, the PDEs are posed on the state space so it is reasonable to think of this error estimate as $O(h)$.

Note that Theorem 5.2 can be proved using the same sequence of calculations and bounds as Lemma 5.3. To see this observe that if we replace $U_{ad,\sigma}$ by U_{ad} in both (5.2) and $(M1_h)$, then u_σ solves the continuous problem (3.1) and $u_{\sigma,h}$ solves the implicitly discretised problem $(M2_h)$.

REMARK 5.1 As we noted in Remark 4.3, sometimes $(M1_h)$ is equivalent to $(M2_h)$. In these cases (e.g. when there are no active control constraints and $V_h \subset U_\sigma$) Theorem 5.2 gives error estimates for $(M1_h)$. This observation proves Corollary 5.1. In particular it gives an estimate for $(M1_h)$ when $n = 3$ without control constraints, which Theorem 5.1 does not provide.

REMARK 5.2 Using this approach to the numerical analysis, the error estimate analogous to Theorem 5.1 for a control problem with an $L^2(\Omega)$ fidelity term (instead of one containing point evaluations) is $O(\sigma + h^2)$ (see Casas & Tröltzsch (2003)).

5.2 Approach 2

This error analysis is based on the technique used in Deckelnick & Hinze (2007) and Leykekhman *et al.* (2013). The approach applies to the implicit discretisation $(M2_h)$ (see (4.13)), and therefore also to the explicit discretisation $(M1_h)$ (see (4.10)) when these discrete problems are equivalent (see Remark 4.3). However it does not apply to $(M1_h)$ in general.

The key ingredient of Approach 2 is bounding the difference between the continuous reduced objective functional applied to the discrete and continuous optimal controls, and similarly for the discrete reduced objective functional. Instead of needing error estimates for the control-to-state operator and its adjoint, which were required to prove Theorem 5.2, this approach only uses the strong supremum norm error estimate (4.7). It also does not require the manipulation of measures. As a result this approach is mathematically simpler than Approach 1, but it still allows us to prove the same error estimate as in Theorem 5.2 (modulo ε).

THEOREM 5.4 Let u be a solution of (3.1) and u_h be a solution of $(M2_h)$ (see (4.13)). Then for any $\varepsilon > 0$,

$$\|u - u_h\|_{L^2(\Omega)} \leq C(\varepsilon) h^{2-\frac{n}{2}-\varepsilon}$$

with C independent of h .

Proof. First observe that

$$\begin{aligned} \hat{f}(u_h) - \hat{f}(u) &= \frac{1}{2} \sum_{\omega \in I} (Su_h - Su)(\omega)^2 + \frac{\nu}{2} \|u_h - u\|_{L^2(\Omega)}^2 \\ &\quad + \sum_{\omega \in I} (Su_h - Su)(Su - g_\omega)(\omega) + \nu(u, u_h - u) \\ &\geq \frac{1}{2} \sum_{\omega \in I} (Su_h - Su)(\omega)^2 + \frac{\nu}{2} \|u_h - u\|_{L^2(\Omega)}^2, \end{aligned} \quad (5.15)$$

since the optimality conditions imply that

$$\sum_{\omega \in I} (Su_h - Su)(Su - g_\omega)(\omega) = a(Su_h - Su, p) = (u_h - u, p) \geq -\nu(u_h - u, u).$$

Similarly

$$\hat{f}_h(u) - \hat{f}_h(u_h) \geq \frac{1}{2} \sum_{\omega \in I} (S_h u_h - S_h u)(\omega)^2 + \frac{\nu}{2} \|u_h - u\|_{L^2(\Omega)}^2. \quad (5.16)$$

Note that the final inequality in this calculation holds for $(M2_h)$ but not for $(M1_h)$ without additional assumptions.

So combining (5.15) and (5.16) we get

$$\begin{aligned} \nu \|u - u_h\|_{L^2(\Omega)}^2 &\leq \hat{f}(u_h) - \hat{f}(u) + \hat{f}_h(u) - \hat{f}_h(u_h) \\ &\leq |\hat{f}(u) - \hat{f}_h(u)| + |\hat{f}(u_h) - \hat{f}_h(u_h)|. \end{aligned} \quad (5.17)$$

We can bound each of the terms on the right hand side of this inequality. Note that

$$\begin{aligned} |\hat{f}(u) - \hat{f}_h(u)| &= \left| \frac{1}{2} \sum_{\omega \in I} (Su(\omega) - g_\omega)^2 - \frac{1}{2} \sum_{\omega \in I} (S_h u(\omega) - g_\omega)^2 \right| \\ &= \left| \frac{1}{2} \sum_{\omega \in I} (Su - S_h u)(Su - g_\omega + S_h u - g_\omega)(\omega) \right| \\ &\leq C \|Su - S_h u\|_\infty (\|Su\|_\infty + \|S_h u\|_\infty + \max_{\omega \in I} |g_\omega|) \\ &\leq C \|Su - S_h u\|_\infty (\|u\|_{L^2(\Omega)} + 1). \end{aligned}$$

So (4.7) gives that for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\begin{aligned} |\hat{f}(u) - \hat{f}_h(u)| &\leq C(q') h^{3-\frac{n}{q'}} \|u\|_{W^{1,q'}(\Omega)} (\|u\|_{L^2(\Omega)} + 1) \\ &\leq C(q') h^{3-\frac{n}{q'}}. \end{aligned} \quad (5.18)$$

In the same way we get that for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\begin{aligned} |\hat{f}(u_h) - \hat{f}_h(u_h)| &\leq C \|Su_h - S_h u_h\|_\infty (\|Su_h\|_\infty + \|S_h u_h\|_\infty + \max_{\omega \in I} |g_\omega|) \\ &\leq C(q') h^{3-\frac{n}{q'}} \|u_h\|_{W^{1,q'}(\Omega)} (\|u_h\|_{L^2(\Omega)} + 1). \end{aligned} \quad (5.19)$$

Since $u_h = \mathbb{P}_{[a,b]}(-\frac{1}{v}p_h)$ we have $\|u_h\|_{W^{1,q'}(\Omega)} \leq C\|p_h\|_{W_0^{1,q'}(\Omega)}$, and the same calculation as in the beginning of Lemma 5.1 gives that

$$\|p_h\|_{W_0^{1,q'}(\Omega)} \leq C(q')$$

independently of h . Combining this, (5.17), (5.18) and (5.19) gives

$$\|u - u_h\|_{L^2(\Omega)}^2 \leq C(q')h^{3-\frac{n}{q'}}.$$

Now for any $\varepsilon > 0$ we can set

$$q' = \frac{n}{n-1+2\varepsilon},$$

which completes the proof of the theorem. \square

REMARK 5.3 In this proof we used the strong supremum norm estimate (4.7) rather than (4.6). This cannot be used to improve the estimates from Approach 1 in Section 5.1; supremum norm estimates are not used in Lemma 5.2, and in Lemma 5.3 we can improve the bound in (5.14) but the error would still be dominated by the $h^{2-\frac{n}{2}}$ term in (5.13).

5.3 Forcing term

We did not include a forcing term in our write up in order to simplify the presentation. However all the results we have proved still hold if we include a forcing term f in the state equation with the regularity $f \in W_0^{1,q'}(\Omega)$ for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$. In particular, if we replace (2.2) by

$$a(y, v) = (\eta + f, v) \quad \forall v \in H_0^1(\Omega), \quad (5.20)$$

and consider a control problem of the form

$$\begin{aligned} \min \quad & J(y, \eta) := \frac{1}{2} \sum_{\omega \in I} (y(\omega) - g_\omega)^2 + \frac{\nu}{2} \|\eta\|_{L^2(\Omega)}^2 \\ \text{over} \quad & C_0(\Omega) \times L^2(\Omega) \\ \text{s.t.} \quad & (5.20) \text{ holds} \\ \text{and} \quad & \eta \in U_{ad} := \{\eta \in L^2(\Omega) : a \leq \eta \leq b \text{ a.e. in } \Omega\} \end{aligned}$$

with all other assumptions the same as in (3.1). This problem has the reduced form

$$\begin{aligned} \min \quad & \hat{J}(\eta) := \frac{1}{2} \sum_{\omega \in I} (S(\eta + f)(\omega) - g_\omega)^2 + \frac{\nu}{2} \|\eta\|_{L^2(\Omega)}^2 \\ \text{over} \quad & \eta \in U_{ad}, \end{aligned} \quad (5.21)$$

where S is as defined previously. For this problem we can construct non-trivial examples with explicitly known solutions (see Section 6.2), which we cannot do for the problem without a forcing term. So after extending our theory to include a forcing term we are able to perform some numerical experiments to verify that our error estimates are observed in practice.

The forcing term means that the mapping from η to y defined by the state equation is no longer linear but instead affine. This difference can be handled with only minor modifications to our problem

formulations and proofs, which we now mention: The optimal control problem with forcing still has a unique solution (see e.g. Theorem 1.45 in Hinze *et al.* (2009)). Corollary 1.3 in Hinze *et al.* (2009) gives that u solves (5.21) if and only if u solves (3.3) with Su replaced by $S(u + f)$ i.e. for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$ there exist $p \in W_0^{1,q'}(\Omega)$ such that

$$\begin{aligned} u &\in U_{ad}, \quad (p - vu, v - u) \geq 0 & \forall v \in U_{ad}, \\ a(v, p) &= \sum_{\omega \in I} (S(u + f)(\omega) - g_\omega)v(\omega) & \forall v \in W_0^{1,q}(\Omega). \end{aligned}$$

The same reasoning applies to the discrete problems and their optimality conditions with the obvious modifications. In particular the optimality conditions for the discrete problem (M2_h) (see (4.13)) with the inclusion of the forcing term are: There exists a $p_h \in V_h$ such that

$$u_h \in U_{ad}, \quad (p_h + vu_h, v - u_h) \geq 0 \quad \forall v \in U_{ad}, \quad (5.22a)$$

$$a(v_h, p_h) = \sum_{\omega \in I} (S_h(u_h + f)(\omega) - g_\omega)v_h(\omega) \quad \forall v_h \in V_h. \quad (5.22b)$$

Theorems 5.1, 5.4 and 5.2 still hold with same methods of proof; the f term slightly alters the calculations but does not cause problems, since it follows immediately from the supremum norm error estimate (4.7) that for $\eta \in W_0^{1,q'}(\Omega)$ with $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\|(S - S_h)(\eta + f)\|_\infty \leq C(q')h^{3-\frac{n}{q'}} \|\eta + f\|_{W_0^{1,q'}(\Omega)}.$$

6. Numerical results

In this section we develop a numerical method for solving (M2_h) with modification to include a forcing term (see (5.22)) and show that the a priori $L^2(\Omega)$ error estimates that we proved for this discrete problem are numerically realised. In order to do this we solve simple examples of the optimal control problems with explicitly known solutions. We also include some simulations for more interesting problems for which the exact solutions are not known.

6.1 Numerical method

We only develop a numerical method for solving (M2_h) because we are able to prove better error estimates for this discrete problem. In particular, we do not have an error estimate for (M1_h) when $n = 3$ with control constraints. Perhaps such an estimate could be proved in other ways, but we cannot easily experimentally investigate if it holds either; we only have explicit solutions (which allow us to reliably test error estimates) when there are no active control constraints. We will now describe the numerical method.

If u_h solves (5.22), then by substituting $u_h = \mathbb{P}_{[a,b]}(-\frac{1}{v}p_h)$ we get that the state $y_h := S_h u_h \in V_h$ and the adjoint variable $p_h \in V_h$ solve

$$\left(\begin{array}{l} a(y_h, v_h) - (-\frac{1}{v}p_h + (a + \frac{1}{v}p_h)^+ - (-\frac{1}{v}p_h - b)^+ - f, v_h) \\ a(w_h, p_h) - \sum_{\omega \in I} (y_h(\omega) - g_\omega)w_h(\omega) \end{array} \right) = 0 \quad (6.1)$$

for all $v_h, w_h \in V_h$. Here v^+ denotes the nonnegative part of v i.e. $\max(0, v)$. Once this problem has been solved, the u_h solving (4.14a) can easily be determined from p_h by setting $u_h = \mathbb{P}_{[a,b]}(-\frac{1}{v}p_h)$. We will now describe a numerical method for solving (6.1) with and without control constraints.

6.1.1 No control constraints. In the case of no control constraints ($b = -a = \infty$) the nonlinear $\max(0, \cdot)$ terms drop out, leaving a linear problem. Let $y_h = \sum_{z \in \mathcal{N}} y_z \varphi_z$ and $p_h = \sum_{z \in \mathcal{N}} p_z \varphi_z$, where φ_z are the usual nodal basis functions of V_h (defined by $\varphi_z(\bar{z}) = \delta_{z\bar{z}}$ for $\bar{z} \in \mathcal{N}$, where $\delta_{z\bar{z}}$ denotes the Kronecker delta and \mathcal{N} is the set of interior vertices of the triangulation), and y_z and p_z are the coefficients corresponding to the basis functions. As we have no control constraints, testing (6.1) with $v_h = \varphi_z$ and $w_h = \varphi_{\bar{z}}$ for all $z, \bar{z} \in \mathcal{N}$ leads to a system of linear equations of real variables. In particular, let \bar{y} and \bar{p} be vectors of coefficients defined by $\bar{y}_z = y_z$ and $\bar{p}_z = p_z$ for $z \in \mathcal{N}$ i.e. use the set of interior vertices as an index. Then we can solve (6.1) by solving the system of linear equations

$$\begin{pmatrix} A & \frac{1}{v}M \\ -\sum_{\omega \in I} M_{\omega} & A \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{p} \end{pmatrix} = \begin{pmatrix} \bar{F} \\ -\sum_{\omega \in I} \bar{G}_{\omega} \end{pmatrix},$$

where matrices A , M and M_{ω} and vectors \bar{F} and \bar{G}_{ω} are defined by

$$\begin{aligned} A_{z\bar{z}} &= a(\varphi_z, \varphi_{\bar{z}}), & M_{z\bar{z}} &= (\varphi_z, \varphi_{\bar{z}}), & (M_{\omega})_{z\bar{z}} &= \varphi_z(\omega) \varphi_{\bar{z}}(\omega) & \forall z, \bar{z} \in \mathcal{N}, \\ \bar{F}_z &= (f, \varphi_z), & (\bar{G}_{\omega})_z &= g_{\omega} \varphi_z(\omega) & & \forall z \in \mathcal{N}. \end{aligned}$$

As the basis functions φ_z are piecewise linear with small support, the integrals that form the elements of the matrices and vectors are straightforward to compute, assuming A and f have a simple form (or else numerical integration of some terms may be required, which we discuss later). The matrix in this system of equations is sparse and so the system can be solved efficiently.

6.1.2 Control constraints. In the case of control constraints the nonlinear $\max(0, \cdot)$ terms mean that we can no longer use the above approach to construct a linear system of equations of real variables. Instead we will solve the problem iteratively using a Newton-type method. Let $F_h : V_h \times V_h \rightarrow V_h^* \times V_h^*$ with $F_h(y_h, p_h)(w_h, v_h)$ defined by the left hand side of (6.1). Then it can be written as

$$F_h(y_h, p_h) = 0 \quad \text{in } V_h^* \times V_h^*. \quad (6.2)$$

The $\max(0, \cdot)$ terms mean that F_h is not Fréchet differentiable. However we can apply a generalised Newton method called the semismooth Newton method (see e.g. Ulbrich (2002) and Hintermüller & Kopacka (2009)). This amounts to applying the Newton method in the usual way but taking the derivative of $\max(0, x)$ to be

$$\max'(0, x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases}$$

So we take an initial guess y_h^0, p_h^0 then apply until the convergence the semismooth Newton iteration

$$\begin{pmatrix} y_h^{n+1} \\ p_h^{n+1} \end{pmatrix} = \begin{pmatrix} y_h^n \\ p_h^n \end{pmatrix} + \begin{pmatrix} \delta y_h \\ \delta p_h \end{pmatrix},$$

where $\delta y_h, \delta p_h$ solve

$$\begin{aligned} & F_h'(y_h^n, p_h^n)(\delta y_h, \delta p_h) \\ &= \begin{pmatrix} a(\delta y_h, \cdot) - \frac{1}{v} \left((-1 + \max'(0, a + \frac{1}{v} p_h^n) + \max'(0, -\frac{1}{v} p_h^n - b)) \delta p_h, \cdot \right) \\ a(\cdot, \delta p_h) - \sum_{\omega \in I} \delta y_h(\omega)(\cdot)(\omega) \end{pmatrix} \\ &= -F_h(y_h^n, p_h^n). \end{aligned} \quad (6.3)$$

Note that if we have no control constraints the first Newton iteration is equivalent to solving (6.1).

As before we can represent δy_h and δp_h as a sum of basis functions weighted by coefficients, and testing (6.3) with the basis functions allows us to construct the following system of linear equations of real variables:

$$\begin{pmatrix} A & \frac{1}{v} M_c \\ -\sum_{\omega \in I} M_\omega & A \end{pmatrix} \begin{pmatrix} \delta \bar{y} \\ \delta \bar{p} \end{pmatrix} = \begin{pmatrix} \bar{R}_1 \\ \bar{R}_2 \end{pmatrix},$$

where

$$(M_c)_{z\bar{z}} := (c(x)\varphi_z, \varphi_{\bar{z}}) \quad \forall z, \bar{z} \in \mathcal{N},$$

$$\text{with } c(x) := 1 - \max'(0, a + \frac{1}{v} p_h^n(x)) - \max'(0, -\frac{1}{v} p_h^n(x) - b),$$

$$(\bar{R}_1)_z := -(F_h(y_h^n, p_h^n)_1, \varphi_z), \quad (\bar{R}_2)_z := -(F_h(y_h^n, p_h^n)_2, \varphi_z) \quad \forall z \in \mathcal{N}.$$

Note that since $p_h^n(x)$ is piecewise linear, the integrals of functions such as $\max'(0, a + \frac{1}{v} p_h^n(x))\varphi_z\varphi_{\bar{z}}$ can be computed exactly. In practice we instead approximate this using a numerical quadrature i.e. instead of $(c(x)\varphi_z, \varphi_{\bar{z}})$ we will compute $Q(c(x)\varphi_z\varphi_{\bar{z}})$, where

$$Q(\eta) := \sum_{T \in \mathcal{T}_h} Q_T(\eta), \quad Q_T(\eta) := \sum_{q=1}^K \hat{w}_q |DF_T(\hat{x}_q)| \eta(F_T(\hat{x}_q)).$$

Here $\{(\hat{w}_q, \hat{x}_q)\}_{q=1}^K$ is a collection of K pairs of weights and points on a reference element \hat{T} and F_T is the reference mapping between \hat{T} and T . We will use a Gaussian quadrature of high order (large K), so $Q(\eta) \approx \int_{\Omega} \eta(x) dx$. We will also use this quadrature rule to approximate f as it may have a form that makes it complicated to integrate by hand. The moderately large error from our discretisation should dominate the smaller error from Gaussian quadrature (as it has good approximation properties), so we do not expect using quadrature to affect the $L^2(\Omega)$ error we observe in practice. Note that using quadrature means that we are not solving (6.1) but rather a close approximation. Although using quadrature is not strictly necessary, the implementation without would require us to do additional calculations by hand, particularly in 3 dimensions. In comparison, there is built in support for numerical quadrature in many finite element software packages.

Define the product space norm for $(z_1, z_2) \in Z \times Z$, where Z is a normed vector space, by $\|(z_1, z_2)\|_Z = \sqrt{\|z_1\|_Z^2 + \|z_2\|_Z^2}$. For $z \in H^{-1}(\Omega)$ let $w \in H_0^1(\Omega)$ be defined by

$$(\nabla w, \nabla v) = \langle z, v \rangle_{H^{-1}(\Omega)} \quad \forall v \in H_0^1(\Omega).$$

Then

$$\begin{aligned} \|z\|_{H^{-1}(\Omega)} &= \sup_{v \in H_0^1(\Omega)} \frac{\langle z, v \rangle_{H^{-1}(\Omega)}}{\|v\|_{H_0^1(\Omega)}} \\ &= \sup_{v \in H_0^1(\Omega)} \frac{(\nabla w, \nabla v)}{\|v\|_{H_0^1(\Omega)}} \\ &= \|w\|_{H_0^1(\Omega)}. \end{aligned}$$

This motivates us to iterate the Newton method until the stopping criterion $\|F_h(y_h, p_h)\|_Z$ is small, where for $z_h \in V_h^*$ we define $\|z_h\|_Z := \|w_h\|_{H_0^1(\Omega)}$ with

$$w_h \in V_h, \quad (\nabla w_h, \nabla v_h) = \langle z_h, v_h \rangle_{V_h^*} \quad \forall v_h \in V_h.$$

Note that if $\|F_h(y_h, p_h)\|_Z = 0$ then (y_h, p_h) is the solution to (6.2). The algorithm we use is stated precisely in Algorithm 1 below.

Algorithm 1 Newton method

Input: T_h, y_h^0, p_h^0 and $\text{DATA} = (\Omega, v, f, a, b, I, \{y_w\}_{w \in I})$ $\triangleright (y_h^0, p_h^0) = (0, 0)$
 1: **while** $\|F_h(y_h^k, p_h^k)\|_Z > \delta$ **do** $\triangleright \delta = 1e-8$
 2: Compute $(\delta y_h, \delta p_h)$ by solving (6.3): $F'_h(y_h^k, p_h^k)(\delta y_h, \delta p_h) = -F_h(y_h^k, p_h^k)$.
 3: $(y_h^{k+1}, p_h^{k+1}) \leftarrow (y_h^k, p_h^k) + (\delta y_h, \delta p_h)$
 4: $k \leftarrow k + 1$
 5: **end while**
 6: **return** y_h^k, p_h^k

Newton type methods typically offer local superlinear convergence. We do not prove this, but we note in Section 6.5 that our algorithm is very effective in practice. On all the problems we tested it provided quadratic mesh independent convergence to the solution even with the bad initial iterate of $(0, 0)$.

6.1.3 Implementation. As we remarked above, in the case of no control constraints the first iteration of the Newton method solves (6.1). So rather than implementing two different numerical methods, we also use Algorithm 1 to solve the problem when there are no control constraints.

We implemented Algorithm 1 in the Distributed and Unified Numerics Environment (DUNE) using DUNE-FEM (see Bastian *et al.* (2008a,b); Dedner *et al.* (2010)). This environment has the advantage that once an algorithm has been implemented, it is straightforward to change features of the implementation that would usually be fixed. For solving the linear systems for each iteration of the Newton method we used the biconjugate gradient stabilised method with an incomplete LU factorisation or Gauss-Seidel preconditioner.

6.2 Exact solutions

We can construct an exact solution for a simple example of the optimal control problem in dimensions 2 and 3 without control constraints. This allows us to verify our error estimates. The key fact we will use to do this is that fundamental solutions of the Laplace equation $-\Delta y = \delta_{x'}$ are given by

$$\begin{cases} -\frac{1}{2\pi} \log |x - x'| + C & n = 2, \\ \frac{1}{4\pi |x - x'|} + C & n = 3. \end{cases}$$

So take $\Omega = B_1(0)$, the open unit ball in \mathbb{R}^n centred at the origin, and $I = \{0\}$. Then

$$p(x) = \begin{cases} -\frac{1}{2\pi} \log |x| (y(0) - g_0) & n = 2 \\ \frac{1}{4\pi} \left(\frac{1}{|x|} - 1\right) (y(0) - g_0) & n = 3 \end{cases}$$

is the unique p solving (3.3b), and $u = -\frac{1}{v}p$ (as we have no control constraints). Note that u and p are unbounded, however they are still $L^2(\Omega)$ functions. To see this note that converting to polar and

spherical coordinates we have

$$\begin{aligned}\int_{\Omega} (\log |x|)^2 dx &= \int_0^{2\pi} \int_0^1 (\log r)^2 r dr d\theta < \infty, \\ \int_{\Omega} \frac{1}{|x|^2} dx &= \int_0^{2\pi} \int_0^{\pi} \int_0^1 \sin \theta dr d\theta d\varphi < \infty.\end{aligned}$$

We can now set y to be any function satisfying the boundary conditions (e.g. $y(x) = \cos(\frac{\pi|x|}{2})$), and take $f = -\Delta y - u$. We also set $v = 1$ and $g_0 = y(0) - 1$ to simplify the problem and exact solution further.

6.3 2D numerical results

Motivated by the above construction take $\Omega = B_1(0)$, $A = -\Delta$, $I = \{0\}$, $g_0 = 0$, $b = -a = \infty$, $v = 1$, and

$$f = \frac{\pi}{4} \left(\frac{2}{|x|} \sin \left(\frac{\pi|x|}{2} \right) + \pi \cos \left(\frac{\pi|x|}{2} \right) \right) - \frac{1}{2\pi} \log |x|.$$

Then the solution to the control problem is

$$\begin{aligned}u(x) &= -p(x) = \frac{1}{2\pi} \log |x|, \\ y(x) &= \cos \left(\frac{\pi|x|}{2} \right).\end{aligned}$$

This solution is interesting because the control is singular (infinite) at the prescribed point $(0,0)$ but it is still an $L^2(\Omega)$ function. We solve this problem numerically using the numerical method outlined in Section 6.1, giving Figure 2. Note that the solution to the discrete problem must be bounded, even though it is approximating an unbounded function. As a result, the magnitude of the spike in u_h notably increases as the triangulation is refined (but $\|u_h\|_{L^2(\Omega)}$ is stable).

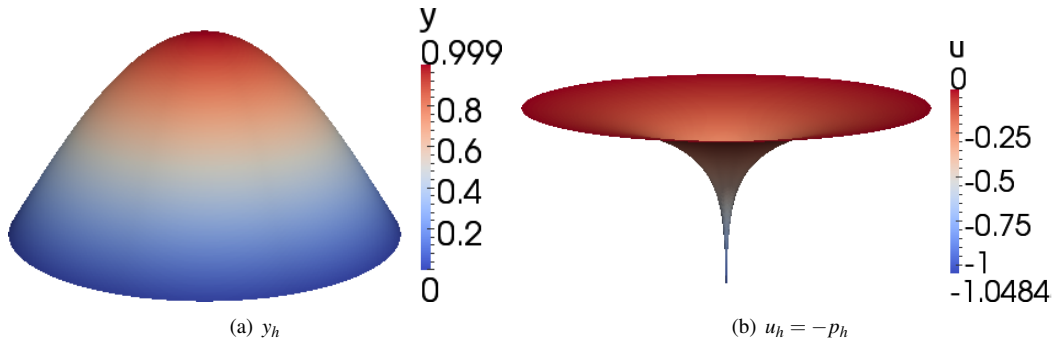


FIG. 2. The radially symmetric solution to our 2D problem with explicitly known solution.

The computed $L^2(\Omega)$ errors are in Table 2, where $\|u - u_h\|_{L^2(\Omega)}$ is approximated using a Gaussian quadrature rule of high order, and the experimental order of convergence is defined by

$$\text{EOC}_h = \frac{\log(\|u - u_{h/2}\|_{L^2(\Omega)} / \|u - u_h\|_{L^2(\Omega)})}{\log 2}.$$

TABLE 2. *EOCs for the 2D problem with explicitly known solution (see Figure 2).*

#DoFs	h	$\ u - u_h\ _{L^2(\Omega)}$	EOC_h
0.5	25	0.03258	-
0.25	81	0.0160362	1.0226543
0.125	289	0.00787259	1.0264221
0.0625	1089	0.00389451	1.0153965
0.03125	4225	0.00193778	1.0070370
0.015625	16641	0.000966977	1.0028513
0.0078125	66049	0.00048313	1.0010701

TABLE 3. *EOCs for the 2D problem on the left hand side of Figure 3, which has control constraints.*

h	#DoFs	$\ u - u_h\ _{L^2(\Omega)}$	EOC_h
0.353553	25	2.8881	-
0.176777	81	1.51039	0.93520339
0.0883883	289	0.80295	0.91153608
0.0441942	1089	0.409627	0.97100093
0.0220971	4225	0.205786	0.99316598
0.0110485	16641	0.100486	1.0341436

The data suggest order h convergence for this problem, which agrees with the estimate we proved in Theorem 5.4

The solution of a more interesting problem including control constraints and more evaluation points is shown on the left hand side of Figure 3. It appears that p_h is approximating an unbounded p , though $\|p_h\|_{L^2(\Omega)}$ is bounded. However u is certainly bounded due to the control constraints. We do not know the exact solution to this problem so we cannot calculate the error exactly. However we can calculate an approximate order of convergence by comparing to the solution on a very fine triangulation i.e. $\tilde{u} = u_{h_{\text{fine}}}$ with $h_{\text{fine}} = 0.00276214$, which corresponds to 263169 DOFs. So we instead compute

$$\text{EOC}_h = \frac{\log(\|\tilde{u} - u_{h/2}\|_{L^2(\Omega)} / \|\tilde{u} - u_h\|_{L^2(\Omega)})}{\log 2} \quad (6.4)$$

for $h \gg h_{\text{fine}}$. We ensure that the fine triangulation is a refinement of the coarser triangulations, so the $L^2(\Omega)$ errors can be computed accurately using a high order Gaussian quadrature. These approximate EOCs can be seen in Table 3. They agree with the error estimate we proved for the case of active control constraints in Theorem 5.4. The slight increase in the EOC for the smallest value of h is expected as we are computing the error against a discrete solution and not the true solution.

On the right hand side of Figure 3 we have the solution of the 2D problem we just considered but without the control constraints. We observe that this allows the state to get slightly closer to the prescribed values. In order to get closer still we would need to decrease v . Figure 4 shows a more interesting example with $v = 1e - 4$ (i.e. very small). As a result the state takes values very close to the prescribed values, and overshoots the value 1 on parts of the domain in order to achieve this.

TABLE 4. *EOCs to our 3D problem with explicitly known solution (see Figure 5).*

h	#DoFs	$\ u - u_h\ _{L^2(\Omega)}$	EOC _{h}
1	27	0.103658	-
0.5	125	0.0719594	0.52657640
0.25	729	0.0474726	0.60008809
0.125	4913	0.0322929	0.55587806
0.0625	35937	0.0225399	0.51873589

6.4 3D numerical results

Similarly take $\Omega = B_1(0)$, $A = -\Delta$, $I = \{0\}$, $g_0 = 0$, $b = -a = \infty$, $v = 1$, and

$$f = \frac{\pi}{4} \left(\frac{4}{|x|} \sin\left(\frac{\pi|x|}{2}\right) + \pi \cos\left(\frac{\pi|x|}{2}\right) \right) + \frac{1}{4\pi} \left(\frac{1}{|x|} - 1 \right).$$

Then the solution to the control problem is

$$\begin{aligned} u(x) &= -p(x) = -\frac{1}{4\pi} \left(\frac{1}{|x|} - 1 \right), \\ y(x) &= \cos\left(\frac{\pi|x|}{2}\right). \end{aligned}$$

This solution can be seen in Figure 5. We observe order \sqrt{h} convergence (see Table 4), which again agrees with the estimate we proved in Theorem 5.4.

6.5 Mesh independence

We finish by justifying the effectiveness of our numerical method. When we have no control constraints the problem is linear and the Newton method always finds the exact solution in a single iteration. When we have control constraints the problem is nonlinear and we still have good mesh independence properties; the number of Newton iterations needed for convergence does not increase as h is decreased. See Table 5 for the number of Newton iterations needed to solve the control constrained example from Figure 3 using the initial iterate $(0, 0)$.

We also observe quadratic convergence of the Newton method on average. See Table 6 for the residuals of the Newton method, again for the control constrained example from Figure 3. In the table

$$\text{EOC}_k := \frac{\log(\delta_{k+1}/\delta_k)}{\log(\delta_k/\delta_{k-1})}, \quad \delta_k := \|F_h(y_h^k, p_h^k)\|_{H^{-1}(\Omega)}. \quad (6.5)$$

References

ADAMS, R. A. & FOURNIER, J. J. F. (2003) *Sobolev spaces*. Pure and Applied Mathematics, vol. 140, second edn. Elsevier.

TABLE 5. *Number of iterations of Newton method.*

h	# iterations
0.0883883	3
0.0441942	3
0.0220971	3
0.0110485	3
0.00552427	3

TABLE 6. *Convergence rate of Newton method.*

Iteration k	$\ F_h(u_h^k)\ _Z$	EOC $_k$
0	0.00285272	0
1	4.38339×10^{-5}	0.85936125
2	1.21172×10^{-6}	2.4577166
3	1.79175×10^{-10}	0

- BASTIAN, P., BLATT, M., DEDNER, A., ENGWER, C., KLÖFKORN, R., OHLBERGER, M. & SANDER, O. (2008a) A generic grid interface for parallel and adaptive scientific computing. Part I: Abstract framework. *Computing*, **82**, 103–119.
- BASTIAN, P., BLATT, M., DEDNER, A., ENGWER, C., KLÖFKORN, R., KORNUBER, R., OHLBERGER, M. & SANDER, O. (2008b) A generic grid interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE. *Computing*, **82**, 121–138.
- BRETT, C., ELLIOTT, C. M., HINTERMÜLLER, M. & LÖBHARD, C. (2013) Mesh adaptivity in optimal control of elliptic variational inequalities with point-tracking of the state. *Interfaces and Free Boundaries (submitted)*.
- BRETT, C. (2014) Optimal control and inverse problems involving point and line functionals and inequality constraints. *Ph.D. thesis*, University of Warwick.
- BRETT, C., DEDNER, A. S. & ELLIOTT, C. M. (2014) Optimal control of elliptic PDEs on surfaces of codimension 1 (preprint).
- CASAS, E. (1985) L2 estimates for the finite element method for the Dirichlet problem with singular data. *Numerische Mathematik*, **47**, 627–632.
- CASAS, E. (1986) Control of an elliptic problem with pointwise state constraints. *SIAM Journal on Control and Optimization*, **24**, 1309–1318.
- CASAS, E., CLASON, C. & KUNISCH, K. (2012) Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM Journal on Control and Optimization*, **50**, 1735–1752.
- CASAS, E. & TRÖLTZSCH, F. (2003) Error estimates for linear-quadratic elliptic control problems. *Analysis and Optimization of Differential Systems*, **121**, 89–100.
- CIARLET, P. G. (1978) *The finite element method for elliptic problems*. Studies in Mathematics and its Applications. North-Holland.
- CROUZEIX, M. & THOMÉE, V. (1987) The stability in L_p and W_p^1 of the L^2 -projection onto finite element function spaces. *Mathematics of Computation*, **48**, 521–532.
- DECKELNICK, K. & HINZE, M. (2007) Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM Journal on Numerical Analysis*, **45**, 1937–1953.

- DEDNER, A., KLÖFKORN, R., NOLTE, M. & OHLBERGER, M. (2010) A generic interface for parallel and adaptive scientific computing: abstraction principles and the DUNE-FEM module. *Computing*, **90**, 165–196.
- GILBARG, D. & TRUDINGER, N. S. (2001) *Elliptic partial differential equations of second order*. Classics in Mathematics, vol. 224. Springer.
- GONG, W., WANG, G. & YAN, N. (2014) Approximations of elliptic optimal control problems with controls acting on a lower dimensional manifold. *SIAM Journal on Control and Optimization*, **52**, 2008–2035.
- GRISVARD, P. (1985) *Elliptic problems in nonsmooth domains*. Monographs and Studies in Mathematics, vol. 24. Pitman Advanced Publishing Program.
- HINTERMÜLLER, M. & KOPACKA, I. (2009) Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. *SIAM Journal on Optimization*, **20**, 868–902.
- HINTERMÜLLER, M. & LAURAIN, A. (2008) Electrical impedance tomography: From topology to shape. *Control and Cybernetics*, **37**, 913–933.
- HINZE, M. (2005) A variational discretization concept in control constrained optimization: The linear-quadratic case. *Computational Optimization and Applications*, **30**, 45–61.
- HINZE, M., PINNAU, R. & ULBRICH, M. (2009) *Optimization with PDE constraints*. Mathematical Modelling: Theory and Applications, vol. 23. Springer.
- LEYKEKHMAN, D., MEIDNER, D. & VEXLER, B. (2013) Optimal error estimates for finite element discretization of elliptic optimal control problems with finitely many pointwise state constraints. *Computational Optimization and Applications*, **55**, 769–802.
- MORREY JR., C. B. (1966) *Multiple integrals in the calculus of variations*. Grundlehren der mathematischen Wissenschaften, vol. 130. Springer.
- PIEPER, K. & VEXLER, B. (2013) A priori error analysis for discretization of sparse elliptic optimal control problems in measure space. *SIAM Journal on Control and Optimization*, **51**, 2788–2808.
- RANNACHER, R. & SCOTT, R. (1982) Some optimal error estimates for piecewise linear finite element approximations. *Mathematics of Computation*, **38**, 437–445.
- RUDIN, W. (1987) *Real and complex analysis*, internatio edn. Tata McGraw-Hill Education.
- SCOTT, R. (1973) Finite element convergence for singular data. *Numerische Mathematik*, **21**, 317–327.
- SCOTT, R. & ZHANG, S. (1990) Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Mathematics of Computation*, **54**, 483–493.
- TRÖLTZSCH, F. (2010) *Optimal control of partial differential equations: Theory, methods and applications*. Graduate Studies in Mathematics, vol. 112. American Mathematical Society.
- ULBRICH, M. (2002) Semismooth Newton methods for operator equations in function spaces. *SIAM Journal on Optimization*, **13**, 805–841.

- UNGER, A. & TRÖLTZSCH, F. (2001) Fast solution of optimal control problems in the selective cooling of steel. *ZAMM Journal of Applied Mathematics and Mechanics*, **81**, 447–456.

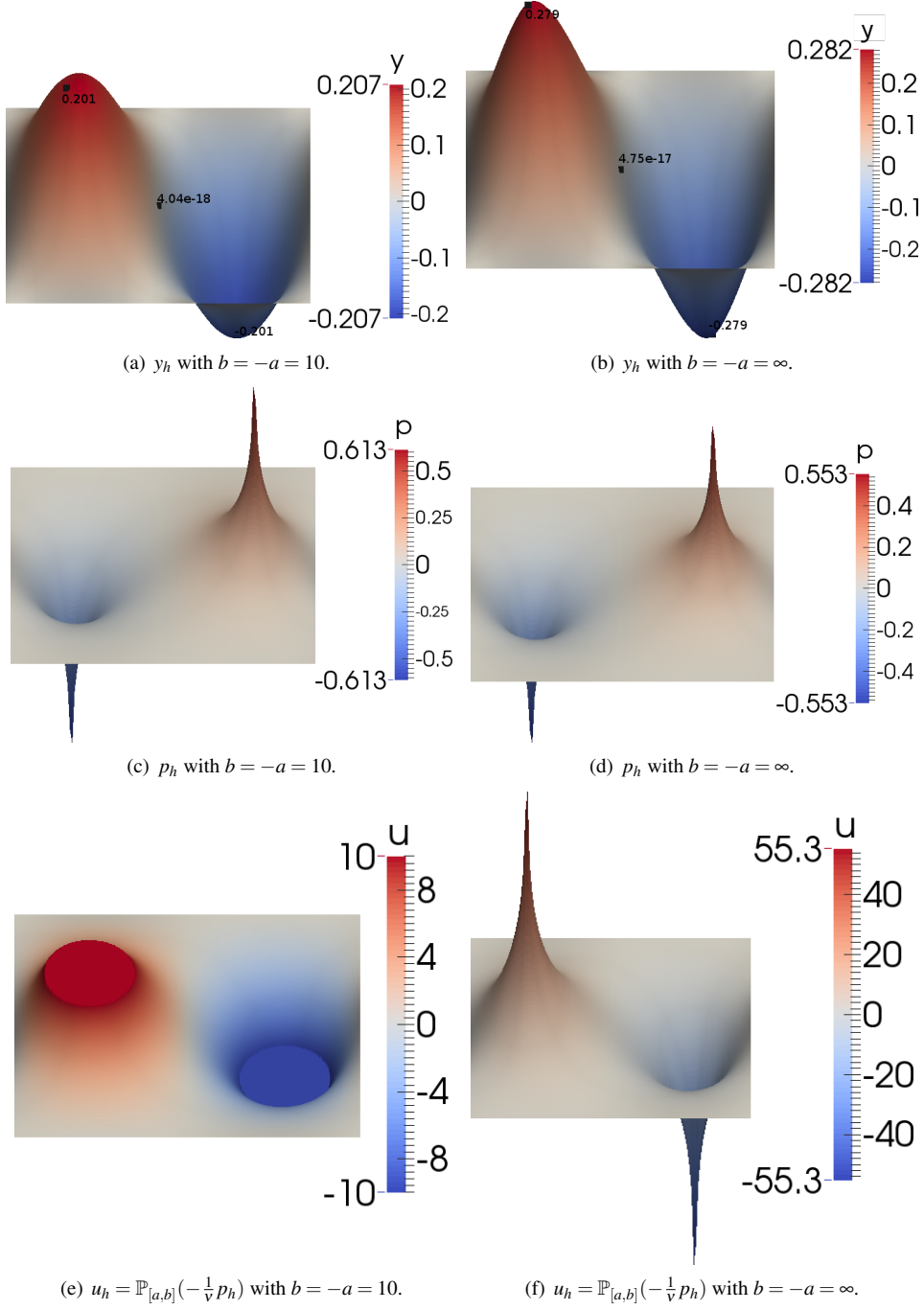


FIG. 3. Solutions for $\Omega = (0, 1)^2$, $A = -\Delta$, $f = 0$, $I = \{(0.2, 0.5), (0.5, 0.5), (0.8, 0.5)\}$, $y_{(0.2, 0.5)} = 1$, $y_{(0.5, 0.5)} = 0$, $y_{(0.8, 0.5)} = -1$, and $v = 1e-2$. The solution on the right has $b = -a = 10$ and the solution on the left has no control constraints ($b = -a = \infty$). The scale on figures that are side by side is the same. The black dots mark the locations of the points in I and the numbers give the value of y_h at these points.

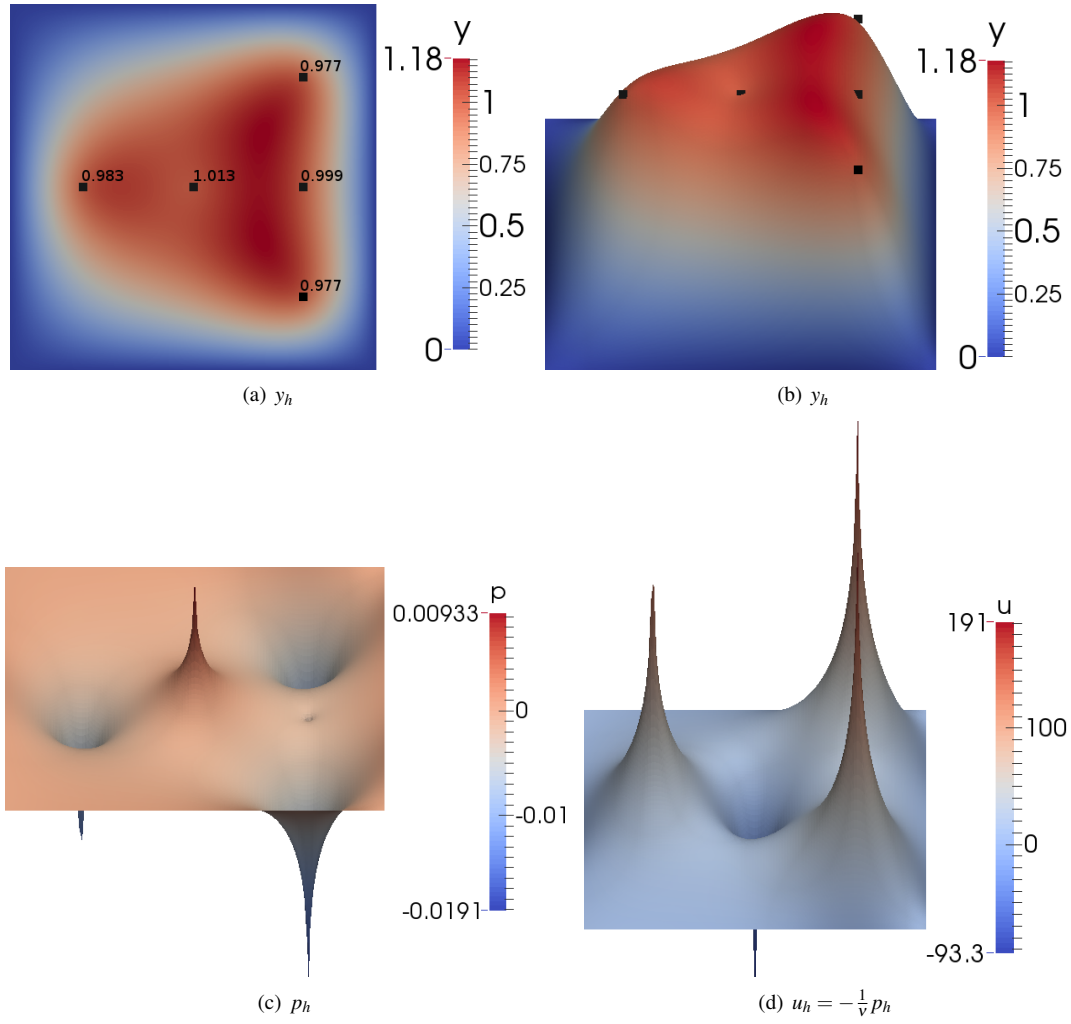


FIG. 4. Solution to a more interesting example with $\Omega = (0, 1)^2$, $A = -\Delta$, $f = 0$, $I = \{(0.2, 0.5), (0.5, 0.5), (0.8, 0.2), (0.8, 0.5), (0.8, 0.8)\}$, $g_\omega = 1$ for all $\omega \in I$, $v = 1e-4$, and $b = -a = \infty$.

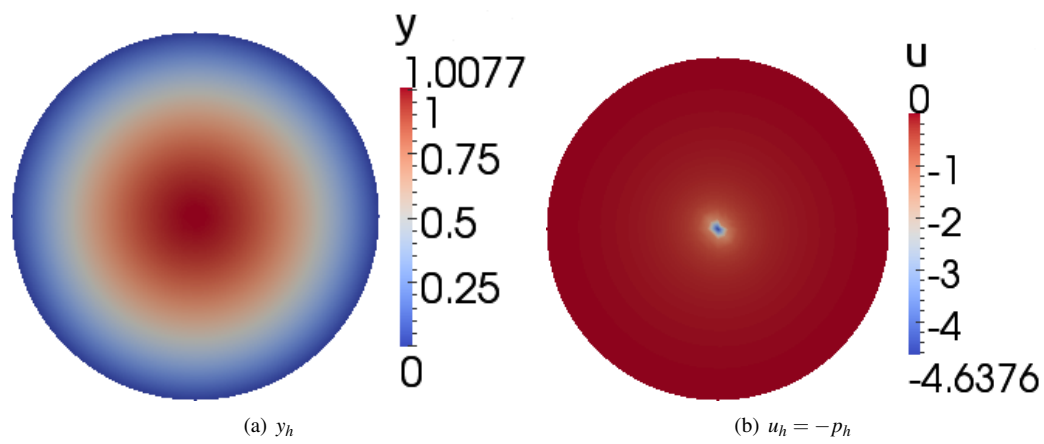


FIG. 5. A slice passing through the origin of the radially symmetric solution to our 3D problem with explicitly known solution.